Integrating visual and tactile robotic perception

Corradi, Tadeo

Award date:
2018

Awarding institution:
University of Bath

Integrating visual and tactile robotic perception

submitted by

Tadeo Mauricio Corradi

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mechanical Engineering

August 2017

Summary

The aim of this project is to enable robots to recognise objects and object categories by combining vision and touch. In this thesis, a novel inexpensive tactile sensor design is presented, together with a complete, probabilistic sensor-fusion model. The potential of the model is demonstrated in four areas: (i) Shape Recognition, where the sensor outperforms its most similar rival, (ii) Single-touch Object Recognition, where state-of-the-art results are produced, (iii) Visuo-tactile object recognition, demonstrating the benefits of multi-sensory object representations, and (iv) Object Classification, which has not been reported in the literature to date. Both the sensor design and the novel database were made available. Tactile data collection is performed by a robot. An extensive analysis of data encodings, data processing, and classification methods is presented. The conclusions reached are: (i) the inexpensive tactile sensor can be used for basic shape and object recognition, (ii) object recognition combining vision and touch in a probabilistic manner provides an improvement in accuracy over either modality alone, (iii) when both vision and touch perform poorly independently, the sensor-fusion model proposed provides faster learning, i.e. fewer training samples are required to achieve similar accuracy, and (iv) such a sensor-fusion model is more accurate than either modality alone when attempting to classify unseen objects, as well as when attempting to recognise individual objects from amongst similar other objects of the same class. (v) The preliminary potential is identified for real-life applications: underwater object classification. (vi) The sensor fusion model provides improvements in classification even for award-winning deep-learning based computer vision models.

Acknowledgements

I would like to thank my supervisors, Dr Pejman Iravani and Prof. Peter M. Hall, for their unrelenting positivity and constructive guidance. I would also like to thank my best friend and wife, Hazel, for her unwavering support throughout.

Contents

List of Figures

List of Tables

List of Acronyms

ABS	Acrylonitrile butadiene styrene
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
GMM	Gaussian Mixture Model
GPU	Graphical Processing Unit
HMP	Hierarchical Matching Pursuit
ICP	Iterative Closest Point
LED	Light Emitting Diode
MEMS	Microelectromechanical Systems
MCA	Maximum Covariance Analysis
μMCA	Mean Maximum Covariance Analysis
MVKDE	Multivariate Kernel Density Estimation
PCA	Principal Component Analysis
RANSAC	Random Sampling Consensus
RGB-D	Red+Green+Blue+Depth
SIFT	Scale Invariant Feature Transform
SOM	Self-Organising Map
SURF	Speed Up Robust Features
SVM	Support Vector Machine
WMCA	Weakly-paired Maximum Covariance Analysis

Chapter 1

Introduction

1.1 Motivation and hypotheses

While it is largely believed that in humans vision is the dominant sense (as exemplified by the Colavita effect [23]), it has been proposed that visual and tactile object representations share information [86]. It has even been suggested that the way in which this information integration is carried out could resemble maximum likelihood integration [32]. It has been noted, however, that the representation of scene layouts in touch and vision are likely not the same, but some form of abstraction is possibly required to make them compatible [87]. So either tactile and visual information are very efficiently combined or they have a shared memory representation [67]. Visual and tactile object representations are intrinsically linked [119]. Multi-modal object recognition achieves view independence easier than either modality alone [68]. The objective of this thesis is to find a fast and robust representation for this integration, which enables a robot to learn about objects from multiple sensors. This will be tested by attempting to perform shape recognition, object recognition and object classification. These problems are well understood in the field of machine vision, but, to date, classification of objects via tactile sensing has not been achieved, neither has it been achieved with the fusion of vision and touch. Humans integrate information from multiple senses to build a representation of the world and in particular of objects. This is necessary since individual senses have significant limitations if taken independently. In particular, in machine vision, some very difficult challenges such as recognising texture, or reflective and translucent objects are significantly easier using touch. In addition, some properties of objects, such as softness or the relationship between

articulate parts, are very difficult to infer from vision alone, and some form of supplementary information may be helpful. The hypotheses of this thesis are:

1. Non-grasping tactile object classification is feasible with a simple, low cost tactile sensor.

2. A simple probabilistic graphical model for the integration of tactile and visual robotic perception is likely to yield higher accuracy object instance recognition and object classification than either modality alone.

Here, *instance recognition* refers to the ability to identify a known object (an object that was present in a training phase, now sensed from a different angle), and *classification* refers to the ability to identify the known class of an unknown object (e.g. a new teddy bear which was not present during training, while other teddy bears were). The hypotheses will be tested by conducting a set of experiments of increasing complexity, from shape recognition to classification. The data sets will be increasingly larger and more challenging.

1.2 Structure and contributions

The work reported in this thesis begins with the design of a new tactile sensor, experiments to find the best way to encode its data leading to tactile shape recognition. Then, an algorithm was designed to use multiple tactile readings from the sensor to recognise a small set of household objects. The algorithm was then extended to incorporate vision (using photos), demonstrating an increase in accuracy. The multi-modal (vision and touch) system was then shown to be able to classify objects within a new (the largest to date) visuo-tactile household object database. Further validation was achieved by showing similar results when the vision model was replaced by a fine-tuned deep-learning award-winning neural net.

This is a *thesis by publication*. Chapters 3-5 are peer-reviewed papers and therefore are enclosed without modification. Chapter 6 is a paper which was submitted to the journal, "Robotics and Autonomous Systems" and is currently under review, and therefore it is included blended into the style of the thesis. Each paper is preceded by a short introduction that summarises and contextualises it within the overarching narrative, providing continuity and cohesion.

Chapter 2 provides an overarching review of relevant background work, including tactile sensors, tactile object recognition, visual recognition and classification and multi-modal fusion. Chapter 3 introduces the novel tactile sensor, its design, data encoding comparisons and basic shape recognition. Chapter 4 describes how tactile data for objects were collected using the robotic arm, and the first experiment pertaining to object recognition. Chapter 5 introduces the sensor-fusion model, along with a comparison to alternatives, for the purpose of object recognition. Chapter 6 introduces the new visuo-tactile database, and the first example of tactile object classification, as well as an example of a potential practical application (underwater object classification). The fusion approach is further validated in this chapter by applying it to a deep-learning vision model. Chapter 7 provides an overall analysis of the results in all contributions, highlighting their strengths, limitations and proposals for further work. It concludes with a summary of the contributions and how they relate to the original hypotheses.

Chapter 2

Literature Review

2.1 Tactile sensing

2.1.1 Tactile sensors

Tactile sensors have been the focus of much research recently [29, 111]. The majority of efforts have been put into low resolution pressure sensor arrays [6, 109, 85, 97, 91, 123]. Many pressure sensor arrays consist of between 4×4 and 32×32 cells (most typically at the lower end of that scale). Each cell is either binary (detects touch or no touch), or pressure sensitive. The low resolution of these sensors means several must be used in conjunction. One of the most widely used sensors, by Weiss[1] has individual cells detecting force and contact and include integrated signal processing to reduce cabling [123]. Attempts have been made to make open source tactile sensors of high reliability and durability and low sensor shift, in order to reduce the cost and increase customisability [56]. This unit, named Takktile TakkArray, is based on arrays of MEMS barometers covered by a rubber membrane, the deformation of which results in the pressure changes being measured. Perhaps the most advanced tactile sensor is the BioTac [124]. The BioTac is capable of sensing pressure distribution via changes in impedance measured at internal electrodes as the internal fluid that resides between them and the deformable outer is displaced due to contact. It is also capable of measuring heat transfer between itself and the contact object, which gives information about the material being touched. A dedicated fluid pressure sensor is used to sense total contact pressure, and the vibrations in this total pressure can be analysed to infer

[1] Weiss tactile sensors. `http://www.weiss-robotics.de/en/english/technology/tactile-sensors.html`

texture when stroking a surface. The BioTac is also one of the most expensive sensors by at least an order of magnitude at present. A popular solution are the tactile sensors built for the icub robot [104]; mounted on the fingertips of the humanoid robot, these comprise 12 capacitive elements, which, when combined, can, for example, estimate the the pressure applied.

Light-based sensors

It is possible to simulate touch by capturing images of the deformation of membranes, as they make contact with an object. Ferrier et al. showed it is possible to reconstruct the shape of a deformable rubber membrane by inspecting the deformation of given patterns, assuming the total energy stored in the configuration would be minimal [37]. Kamiyama et al. [60] used two colours of markers at different depths within an elastic translucent layer to analyse the relative deformation between layers to infer magnitude and direction of applied forces. The OptoForce sensor [2] uses four light emitters and a single receptor, enclosed in a semispherical rubber membrane coated internally with a reflective layer. By detecting the reflexion of light, they are able to very finely infer pressure in three axes. Knoop et al. [63] use an opaque rubber membrane, internally painted with semi-randomly placed white dots, and a low-resolution, high-frequency camera to capture the location and track the movement of these dots. The sensor can run in two modes: high-frequency 'Reflex' mode, where statistics such as displacement are calculated by the on-board sensor circuitry, helpful for detection of contact, and lower frequency 'Explore' mode, where the full 32-by-32 image is transmitted an analysed externally, for example for the estimation of precise forces.

One of the closest related sensor to the BathTip is the GelSight [58], which uses multiple light sources and frequencies, and a high-resolution camera, to capture the deformation of an elastomer, and is able to reconstruct 3D surfaces to microscopic precision. The similarity between this sensor and the BathTip stems from the idea of capturing the deformation of a deformable membrane. The purpose of the elastomer in the GelSight sensor is to conform closely to the shape of the surface being touched, so as to effectively reproduce the surface with a coating whose reflective properties are well known. In contrast, the sensor presented in this thesis makes no such assumption. The shape of the deformed rubber membrane is meant to only loosely respond to large scale properties of

[2] https://optoforce.com/

the surface in contact.

The sensor that inspired the creation of the BathTip is The TacTip (Tactile fingerTip) [20]. It is a biologically inspired tactile sensor based on the deformation of a silicone rubber hemispherical surface and the consequential displacement of a number of internal papillae (internally portruding antennae, whose tips are painted white, so as to exacerbate the deformation of the membrane). A digital camera is used to observe this displacement. This sensor was shown to achieve a high degree of accuracy in sensing edges [21], showing small objects are often clearly identifiable by a human from its tactile image. It has been theoretically shown to have potential in tele-surgery [98]. More recently it has also been successfully used to identify textures [126] by analysing the vibration of a central papilla, and also in reconstructing 2D shapes from the autonomous exploration and feature extraction [5]. It is remarkable that such a simple design can achieve localisation resolutions of the order of 0.1mmm [71], with potential already demonstrated for quality control in production lines [72]. More recently, the TacTip was proven useful in maintaining control of a rolling cylinder, even after forced perturbations [27]. The presence of papillae markers inside the sensor are the key to many of these achievements, as they allow for a simple approach for the measurement of vibration, shear, and torsion. Neither the GelSight nor the BathTip have this capability, and, in fact, some of the recent experiments with the GelSight, demonstrating capability to detect slip [128] and estimation of hardness [130] use a new version of the sensor that includes additional internal markers.

Samples of some of the tactile sensor mentioned in this section can be seen in Fig. 2-1.

2.1.2 Tactile information encoding

It has recently been suggested that basic encodings such as edge orientation and even edge displacement are detected by humans in first order tactile neurons, not in the cerebral cortex [94]. Tactile sensing in robots is often comprised of many tactile arrays which give rise to too much information to handle directly. Some form of Tactile information encoding is therefore desirable.

Finding the best way to encode tactile information is an open problem, and it is strongly dependent on the sensor used. Often a tactile sensory signal corresponds to a heat map of pressure forces, so it makes sense to talk about a “tactile

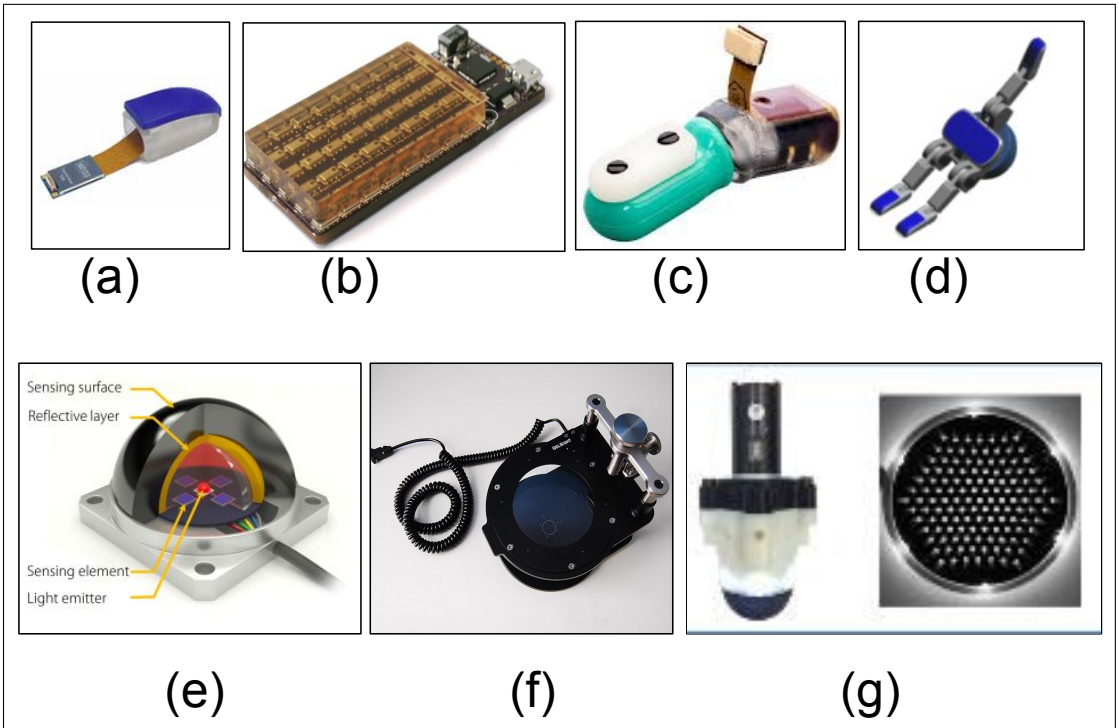


Figure 2-1: *Most widely used tactile sensors: (a) Weiss sensor [123], (b) Takktile sensor array [117], (c) BioTac [124], (d) PPS sensors on a gripper (http://www.pressureprofile.com/), (e) OptoForce (optoforce.com), (f) GelSight sensor [58], and (g) a recent version of the TacTip [20]*

image". The simplest approach is to use these tactile images (pressure images or binary touch images) directly with no encoding and use a simple distance metric [105] to compare similarity. Pattern recognition techniques have been used to find the best encoding function automatically. Early approaches focused simple Artificial Neural Networks (ANNs) [115]. One limitation of ANNs is that they require a large amount of data to be trained. This is at odds with tactile information gathering, which at present is slow due to the robotic manipulation limitation (i.e. the need to make contact with an object means considerations about robot planning must be made). Self-Organizing Maps (SOMs) [64] have been adapted for the purpose of fusing proprioceptive and tactile input for object recognition [59, 97].

Alternatively, preprocessing this tactile information can be done by hand-crafted features. For example, it has been shown that with sufficiently many sensors, extracting image intensity linear moments and extrema is enough to perform object recognition to high accuracy [109]. The performance of Principal Component Analysis (PCA), moment analysis and binary images have been compared in haptic object recognition [45], concluding that central moments outperforms the others. It is also possible to combine the two approaches, for example using PCA and SOMs to extract tactile features which were then used for object recognition [85].

A thorough overview of in tactile sensor technologies in general can be found in [29].

2.2 Tactile object recognition

A large amount of effort has been put into texture recognition [74, 107, 55, 30], since texture is usually difficult to capture from vision alone. This usually involves performing frequency analysis on the vibrations of an end effector which scratches the surface in question. Other common applications of tactile sensors include object localisation [89], slippage detection [17], and grasp stabilisation [8]. Even if the object itself is not labelled, it is possible to extract important information about it using tactile information, such as location and pose for the purposes of grasping and manipulation [90], softness of material [88] and internal states such as whether a bottle is full or empty [18]. Object recognition is a well-studied topic in Computer Vision, however tactile approaches are still relatively few in number.

2.2.1 Volumetric representations

One possible approach to object modelling involves encoding information about its geometry by means of a volumetric representation. Polyhedral approximations have been used successfully, albeit with a limited number of basic geometric solids [16]. Another possible approach is to subdivide the work area into voxels (discrete subdivision of space) and model knowledge about each voxel, which can be used to perform intelligent exploration, even considering empty space. Such an approach combined with primitive tactile feature matching has been applied successfully using Iterative Closest Point (ICP), an iterative procedure which converges to a local optimum match between sets of points for the model and the data, for object recognition [44].

More commonly, however, point-clouds of contact points are used. Point clouds have the advantage of being easily integrated into vision [46]. One clear disadvantage is that, usually, point-clouds comprise many data points and object recognition using direct point cloud comparison is not possible in real time. They are also susceptible to errors and are not robust to changes in the environment during sensing. Work by Bierbaum et al. [11] shows how a point cloud is obtained via exploration performed by a human wearing a data glove. They later refined an algorithm for estimating the shape to be recovered using superquadrics [10] yet their system was not tested on a real robot. Point clouds including surface normal information have been used to reconstruct solid shapes using simulated tactile sensing [46]. In that system, touches from a simulated anthropomorphic hand are used to extract points of contact and surface normals. These points are then used to generate feature vectors that describe the objects, including basic object properties such as proportions, dimensions, and histograms of orientation of surface normals. Crucially, they compare between spherical models and voxel models. Spherical models are histograms of oriented normals mapped onto spherical coordinates. Voxel models simply count the number of points over a 3D grid subdivision, after some regularisation, and apply PCA to reduce dimensionality. The spherical model obtains the best recognition accuracy (93% with a feature vector of size 10) for a set of 15 objects using 375 contact points for training and 375 for testing. While some translation invariance is obtained, the method is susceptible to missing information, such as unreachable parts of the object. A combination of features and volumetric is used by [1]. They employ Random Sample Consensus (RANSAC) [39] to evaluate matches between

sets of tactile features to a number of running hypotheses. They also maintain a voxel representation of the work space to keep track of empty space. Finally, it is merged with a form of point cloud and ICP for hypothesis verification. The system achieves 80% recognition with fewer than 10 touches, from a database of 45 objects. Object models need to be known in advance, which in their context (deep sea object recognition) is not a major limitation.

One way to solve the problem of having too much data (especially potentially redundant data) is to merge points that are close into a probability point modelled by a Kalman filter. This was proven to be achievable in real time and with no significant error with respect to a direct ICP [83]. Attempting to address the sparsity and noise problems in point clouds obtained by tactile exploration, Jin et al. [57] use clustering to subdivide the point cloud into regions which are then encoded as features. These features are then classified using a Gaussian Process (with a squared exponential kernel) for object classification. This is therefore a bag-of-tactile-features approach (see section 2.2.3). Simulations of 8 shape primitives (e.g. pyramid, cone, etc.) give a high accuracy for recognition.

2.2.2 Recognition by grasping

More recently, there have been several projects involving recognition by grasping using machine learning techniques. PCA, SOMs and ANNs have been combined to process the output of Weiss tactile sensory arrays attached to a number of robotic end effectors, to classify household objects [85]. Novel recursive Gaussian kernels have been designed to encode the various stages of contact during grasping leading to a robust on-line system capable of learning new models and classifying objects in real time [109]. In the field of on-line spatio-temporal unsupervised feature learning for object recognition from grasping, the best results are currently obtained by [81]. They extend Hierarchical Matching Pursuit (HMP, a multi-layer hierarchical feature learning system) to include temporal information. They test their method on 6 tactile databases and produce an accuracy of between 80 and 100%. One of the advantages of grasping is that pose ignorance is not that important, since the grasping action can often result in the object coming to one of a small number of possible poses [116]. Using the most advanced tactile sensors (BioTac) and grasping, it has been shown possible to distinguish between 49 objects with almost perfect accuracy [53].

Grasping, and therefore combining proprioceptive, pose and tactile informa-

tion is likely to yield better results than either modality alone [62, 45]. Using grasp, however, limits the size of the object to be identified, requires a robotic hand, and requires a grasp to be achieved.

2.2.3 Bag of tactile words

Bag-of-visual-words models are models that use local feature information to describe an object [28], ignoring relative position between these features and any other global properties of the object. They were originally used for document classification based purely on word counts and not word location or global document structure, hence the name. Similarly, "Bag of tactile words" approaches, here, refer to those that use local tactile information in parts of the object but disregard their geometric location, and any other global object properties such as dimensions, pose or location. These approaches are therefore robust, in principle, to changes in the aforementioned parameters. Discarding those data, however, can be the major limitation since crucial information is lost.

One of the first attempts at a tactile-only object recognition is given by [101]. They use geometric features such as lines and points, together with their evolution over time (named tactemes, since they are similar in concept to phonemes in speech recognition). Their accuracy recognising objects is high (83%); however, the number of shapes is only 6 and they are very basic predefined geometric solids (cylinder, cone, etc.).

Schneider et al. [105] use repeated application of a two fingered grasps using a gripper equipped with Weiss tactile array sensors. Features are extracted, then a bag-of-features approach is used to recognise household and industrial objects. They provide an interesting information theoretic approach for maximum expected information gain to inform grasping position. Using histogram intersection [52] as a measure of similarity, they obtain an accuracy of 84.6% in recognition. They use 830 tactile images for training and 8 to 10 grasp actions to achieve this accuracy, which equates to 16 to 20 tactile images in the testing set. The object pose is strictly known and unchanging (small translation variance is tolerated). It could be argued that this work uses proprioception (they know the height of the gripper) and thus is not purely bag-of-words.

Pezzementi et al. [91] use simulations to compare various methods of feature extraction, and create clusters of these features to compile feature histograms to be compared (using Kullback-Leibler (KL) divergence [66] minimisation) for

recognition. Out of all the feature extractions they tested, after 10 samples for testing, the best performance on simulations was given by Moment-Normalize (65%), whilst on real physical experiments was given by Polar Fourier (70%) (different features). The real testing was done using DigitTacts sensors over a set of 5 objects (the context was recognition of plastic letters) using a basic top-down approach for sensor readings, giving a baseline chance accuracy of 20%. The accuracy of the system improves significantly if more samples are provided for testing, particularly in simulations. It would be interesting to see this system tested on a real scenario with a larger and more varied set of objects, since its simulated performance is promising. The learning phase required 384 tactile images per object.

Luo et al. [80] use a Weiss tactile sensor mounted on a robotic arm to explore and build models for 10 objects, using an adaptation of the SIFT descriptors for the tactile images, removing scale hierarchy and location, as they are redundant for tactile sensing. They discard pose information and thus build a pose invariant model. In their initial model, they use a dictionary learning stage whose dimensionality they optimise to 50. In their subsequent work [78], a novel semi-supervised method is presented, whereby the dimensionality of the dictionary is optimised automatically, further increasing recognition accuracy in a larger (12) and more challenging (higher similarity between items) data set.

Regoli et al. (cite Controlled Tactile Exploration and Haptic Object Recognition, as yet unpublished) achieve an outstanding tactile-only object recognition accuracy (99%) in a data set of 30 objects (some of which are quite similar), by means of stabilising grasping, and performing tactile exploratory procedures. This is particularly impressive given the simplicity of the approach: they use a form of least squares classification on the vector resulting from the concatenation of pressure response vectors.

2.3 Visual object recognition and classification

The field of visual object recognition/classification is large and a detailed analysis goes beyond the scope of this thesis. This section covers main approaches. A detailed review can be found elsewhere [114].

2.3.1 Object recognition: early approaches

The first problem (recognition) relates to the need to identify a known object, perhaps viewed from a novel viewpoint or under different conditions (e.g. lighting, occlusion). Early approaches were based on geometric properties, polyhedral simplification (assuming objects are fully or partly made of polyhedra), generalised cylinders (an attempt to account for non-flat surfaces), aspect graphs (sets of 2D views linked by a graph representing their relative position), feature matching (searching for key local features and their relative position after an affine transformation), and appearance methods (considering the full image, and performing dimensionality reduction). A detailed survey of early approaches can be found in [84].

2.3.2 Object recognition: feature-based approaches

Since the turn of the century, approaches based on local features have gained traction [77, 99, 43]. The idea is to focus on a small region of an image and to encode it using a transformation that will remain largely constant if said part is photographed under different conditions (e.g. scale, lighting, angle). Recognition is then performed by attempting to match these features to known images, either by alignment, considering the relative position of features (e.g. [99]), or by description only, ignoring the features' absolute and relative locations (e.g. [28]). The latter are commonly referred to as 'bag of features' approaches.

2.3.3 Object classification

Object classification, by contrast, is a much more difficult problem. It aims to create a higher level of abstraction: a model for generic object categories. If a new *object instance* of a known category is sensed, it should be assigned to that category, even if the object itself had never been seen before. The dominating approaches for classification are part-based models, bag-of-features, and deep learning.

Part-based models

Originally proposed by [38], part-based models refers to the family of approaches that aims to represent an object class by its structure, i.e. the relative location of its parts. Parts are image patches considered similar based on their appearance.

The problems of deciding what region of an image is a part, and identifying the class of the image are generally solved simultaneously. If each part is considered as the node in a graph, and edges represent the importance and relationship between parts, assumptions can be made about the type of graph that can emerge. Assuming all inter-node relationships are important results in a fully connected graph in what is called constellation models [15]. If a central part is assumed and only relative positions to this part are considered, the resulting graph is a tree [34]. Typically, a custom designed energy function is minimised which jointly penalises matched parts appearances and overall structure dissimilarity between candidate classes and a test image. If no part interrelationships are considered, and only their appearance is compared, the method collapses to bag-of-words (See Section 2.2.3, also below). Part-based models have the advantage of encoding structural information about the object class as well as localised appearance information. This makes them robust to variations in appearance and occlusions. It also makes them particularly suited for object detection (where in the image is a given object, if at all), as matching structure results in part (and therefore object) location. Key disadvantages include the difficulty of matching graphs and the computational complexity of the joint energy minimisation.

An overview of part-based models is given by [35]

Bag-of features models

Bag-of-features approaches in general were discussed in Section 2.2.3. Specifically for vision, the seminal work was done by Csurka el at. [28], comparing simple classifiers on histograms of SIFT features. The pipeline is similar to the process described in Section 2.2.3. First, a visual 'vocabulary' is formed by clustering a large number of feature descriptors. The clusters then act as 'visual words' (following the equivalence with document classification), hence the name. Images are processed by extracting features and using a proximity measure to assign the closest 'visual word', and are thereafter represented by a histogram (or bag) of such words. This quantised approach is not universal, however, as it is possible to perform similarity comparison between vectors of different sizes [47]. The main limitations of bag-of-feature approaches are the lack of relative spatial information and the strong dependency on the choice of visual feature. In particular, feature detectors may be unsuitable for a given class. This can be ameliorated by using dense features (extracting features throughout the image, not just where the

detectors identify a point of interest) [33]. However that brings complications with respect to performance and storage, and the debate on whether dense or sparse features are to be preferred is not settled.

An extended review of bag-of-feature approaches is given by [133], including a comparison between a range of descriptors and classifiers.

Deep Learning

The most recent (and arguably the most successful [50]) approaches to object classification have been based on neural networks of increasing depth and a large number of parameters, they are collectively referred to as Deep Learning approaches [69]. Previously mentioned approaches attempt to define a specific visual feature, or aim to prescribe the potential variables to consider to extract and compare structure, and perform classification on the resulting vectors or representations. Deep learning relies on end-to-end classification, where images are used directly as input, and object class is used as output. It is left to the neural net to 'discover' suitable low-level encodings and useful higher level abstractions to maximise performance. Perhaps the best known deep nets are convolutional neural nets (CNNs, originally introduced by [40]), which force local patches of an image to be treated equally by the neural net classifier, no matter their location, thus enforcing position invariance, and reducing the number of parameters at the same time.

An overview of deep learning methods is given by [103].

2.4 Visuo-tactile integration

Early attempts used vision to guide a series of exploratory behaviours and combined vision and touch to create a modular geometry-based model of an object [2]. Once possible matchings are identified, the robot proceeds to verify a hypothesis by sensing parts of the object which are yet unseen. Rafla, in their PhD thesis [95] developed a method to integrate tactile and (virtual) visual range data to recreate surface equations analytically and perform recognition on simple objects. Their work focuses on surface normals and does not delve into more complex tactile or 3D visual features. Haptic and vision sensors have been used to estimate parameters of a kinematic model for hand-object interactions [4]. Other early efforts have gone into integrating vision, touch, heat and vibration sensors us-

ing a multi-layer ANN [115]. Their system is capable of classifying 14 objects; it is at times perfect in accuracy, yet one must bear in mind that robot-object interactions are pre-programmed, neural nets are trained independently for each modality, and images are taken and classified in advance. Integration by use of direct ANNs has also been shown to be effective at recognising 11 objects when the tactile information is simply the reaction force of robot fingers during grasping [62], clearly demonstrating that accuracy improves when both modalities are considered, over either modality alone. Integration of modalities has proven very valuable in pose estimation for manipulation [93], using a hierarchical approach, where vision and touch are graded for their reliability and preferred accordingly. Proprioception has been combined with vision to perform pose estimation of grasped objects, using simple vector concatenation and an extended Kalman filter [51]. Guler et al. [49] combine vision and touch to determine the content of a number of containers, by grasping, squeezing, and observing the results. They conclude that a multi-modal approach is superior to either sensor alone, as it provides complementary information.

Another advanced system for sensor fusion aims to learn weak pairings between modalities [65]. Their two methods are based on Maximum Covariance Analysis (MCA)[118] (a tool for dimensionality reduction of paired data). The first, called Mean MCA (μMCA), performs strong direct pairing between the mean value of various readings, and therefore it is robust to having many readings from vision and very few from touch. The second, called weakly paired MCA (WMCA)[65], allows for any pairing to be formed between modalities, restricted to pairings within defined groups (so the matching matrix is in block diagonal form), and optimises the choice of pairings. One particular advantage of this approach is the fact that both modalities are only needed during training, so if either is not present during classification, the system will still perform well. In fact, they show that performance on single modality classification is better if both modalities are used in training, so the system does not get "confused" by the additional data, but instead can use it to create a more robust internal representation of sensory information. The application domain used is material recognition, which is performed in a strongly controlled environment. There is potential for a similar approach to achieve tactile-visual feature matching.

Recent work employs an extended Kalman filter to build a refine 3D models of the sensed objects fusing sensory information from an RGB-D camera and a set of binary tactile sensors in a robotic hand which is grasping the object [54].

In this case, the sensor fusion is performed via expectation maximisation on the probability that each cloud point corresponds to a certain tactile point. One of the benefits of this approach is that it is robust to perturbations in the object location due to the grasping procedure, as object velocity is encoded as part of the Kalman Filter. Their main focus is the accurate 3D reconstruction of the objects instead of recognition. Still, the increased accuracy in the reconstruction could then be used for point-cloud based object recognition using methods such as local point-cloud cluster descriptors [42]. A major limitation of this approach, however, is that it makes two strong assumptions: objects are symmetric and are only perturbed on a plane perpendicular to the supporting plane.

2.5 Visuo-tactile object recognition

Recognition of objects using a combination of vision and touch was demonstrated by Yang et al. [127], and Liu et al. [75].

In [127], vision and touch are combined by the concatenation of feature vectors. A simple weighted nearest-neighbour classifier is used, where the weights attributed to vision and touch are a parameter to be learnt and optimised. Their solution is able to recognise any one of 18 household objects, some of which are very similar (such as identically shaped bottles, which differ only by their visual plastic labels) with reported accuracies of over 90 per cent. In all of their experiments, the sensor fusion model outperformed either modality alone.

In [75], a novel sparse coding algorithm is presented to attempt to detect weak pairings between the tactile vector (directly extracted from the sensor) and the visual vector (a covariance matrix of feature descriptors at various windows within an object's photo). This approach seems to outperform the earlier nearest-neighbour counterpart [127]. The 18 objects could be subdivided into 5 classes, and while explicit classification was not attempted, the confusion matrices reveal that most of the uncertainty arose within-class, so the potential for classification is established.

2.6 Tactile and visuo-tactile object classification

To this date, tactile-only and visuo-tactile object classification (recognising the known class of an unknown object) has not been achieved. Multi-sensory object

representations are gaining traction in the literature [61]. The first large visuo-tactile database could soon be a reality [14]. A form of tactile classification is classifying object by attributing binary adjectives (e.g. soft, coarse) using touch [22], which was also attempted using deep learning, comprising of two deep layers, one for vision and one for touch, finally connected by a fusion layer [41]. Deep learning nets of a similar topology were also used by Zheng et al. [134] to classify textures, The work of Sanchez-Fibla et al. [102] hints at the potential for classification using curvature prediction using vision and touch. Tactile-only shape recognition of a small set of shape primitives (cone, cuboid, cylinder, ball, prism) was performed by [48]. Since the given shape models are learnt during training, this can be seen as a form of classification, with prescribed distinct shapes.

Perhaps the most similar work to the one presented in this thesis can be found in [129]. There, 118 fabrics are photographed and 3D scanned while draping from a platform and are also touched with a GelSight sensor (placed on a flat surface, laying flat and folded). The project mainly focuses on performing joint learning from multiple modalities using Convolutional Neural Networks (CNNs), in such a way that the learnt embeddings are similar for similar fabrics. What is remarkable is that the neural nets trained on multiple modalities produce embeddings that allow better matching even using vision alone. By contrast, in this thesis, a related but different problem is tackled, investigating how to learn object categories for a more varied set of objects (household objects), where the tactile perception is likely to be significantly different for different readings. Contact sensations for a fabric laying flat are likely to be similar for various readings. For objects such as shoes or bottles, the tactile sensations will vary greatly depending on the contact location. The aim is also to only loosely control the data collection, to be performed by a robot in a random fashion, simulating some of the conditions of a robot exploring an unknown object.

Chapter 3

Shape recognition with a novel inexpensive tactile sensor

3.1 Motivation: deciding to create a new sensor

A thorough review of available tactile sensors was performed early (see Section 2.1.1), aiming to choose a suitable device for the purpose of this study. Most commercially available sensors proved either too low resolution, too expensive, or both. Two sensors were within the budget of the project: the TacTip [20] and the Takktile [56]. Of the two, the TacTip was chosen as it has higher resolution, it was readily available (a loan of a prototype was secured from the Bristol Robotics Laboratory) and the designers were intending to make the sensor available open-source. Preliminary tests were conducted with this prototype. By the end of the lease, extensive attempts were made to recreate the TacTip (see Section 2.1.1) here at Bath. A number of different rubber membranes were cast, which initially lacked the internal papillae due to insufficient resolution of the 3D printers used for the cast. The idea of using smooth membranes was a result of this complication.

3.2 Summary: sensor design

The sensor is inspired in the working principle of the TacTip (see Section 2.1.1). Simplifications in the TacTip design were identified so that the new sensor would have no papillae nor internal gel. Instead, it has a plain black matt smooth opaque silicone rubber hemispherical membrane, mounted at the end of a rigid

opaque encasing for the digital camera. The body of the sensor was designed to fit the chosen USB camera and commercially available rubber membrane. The length was chosen to allow for the field of view of the camera to capture the entire membrane. The camera incorporates a set of white LEDs which illuminate the rubber membrane from within. When the sensor is in contact with an object, the shading pattern on the membrane changes accordingly and is captured by the camera. Figure 1 in the paper shows the sensor design and a diagram of its functioning principle.

The membrane is 1mm thick, its internal diameter is 40mm, and its external diameter is 42mm. Its main purpose is to render sensing invariant to light conditions and colour.

The encasing has a cylindrical top designed to fit the membrane about it. Its base is squared, designed to securely fit an off-the-shelf e-secure digital USB camera[1], and has a groove to allow space for the camera cable. Its length is the minimum needed to allow the camera to capture the full membrane in its field of view. The encasing was made using 1.75mm Acrylonitrile butadiene styrene (ABS) plastic with a UP3D Plus 3D printer. The camera itself has a resolution of 640 by 480 pixels, at 30 frames per second. The 3D file used to print the encasing (STL file), as well as the Freecad[2] model and links to the other parts are available online[3].

Isometric views of the sensor and detailed schematics can be found in Figure (3-1).

3.3 Results: sensor evaluation

In order to test whether the simplified sensor was suitable for the overarching aim of this project, the first experiment designed involved basic shape recognition. The first step was to find a robust combination of a low-dimensional representation of tactile images and choice of classifier. Various systems were tested in their ability to accurately distinguish between a small number of tactile shapes: nothing, corner, edge, point, curved, spherical, flat, flat-to-edge. The same experiments were run using the TacTip to provide a comparison. The new sensor outperforms the TacTip at shape recognition. This may be due to the fact

[1] https://goo.gl/KseVHG
[2] https://www.freecadweb.org/
[3] 3D model of the tactile sensor encasing available at: `https://github.com/Exhor/bathtip`

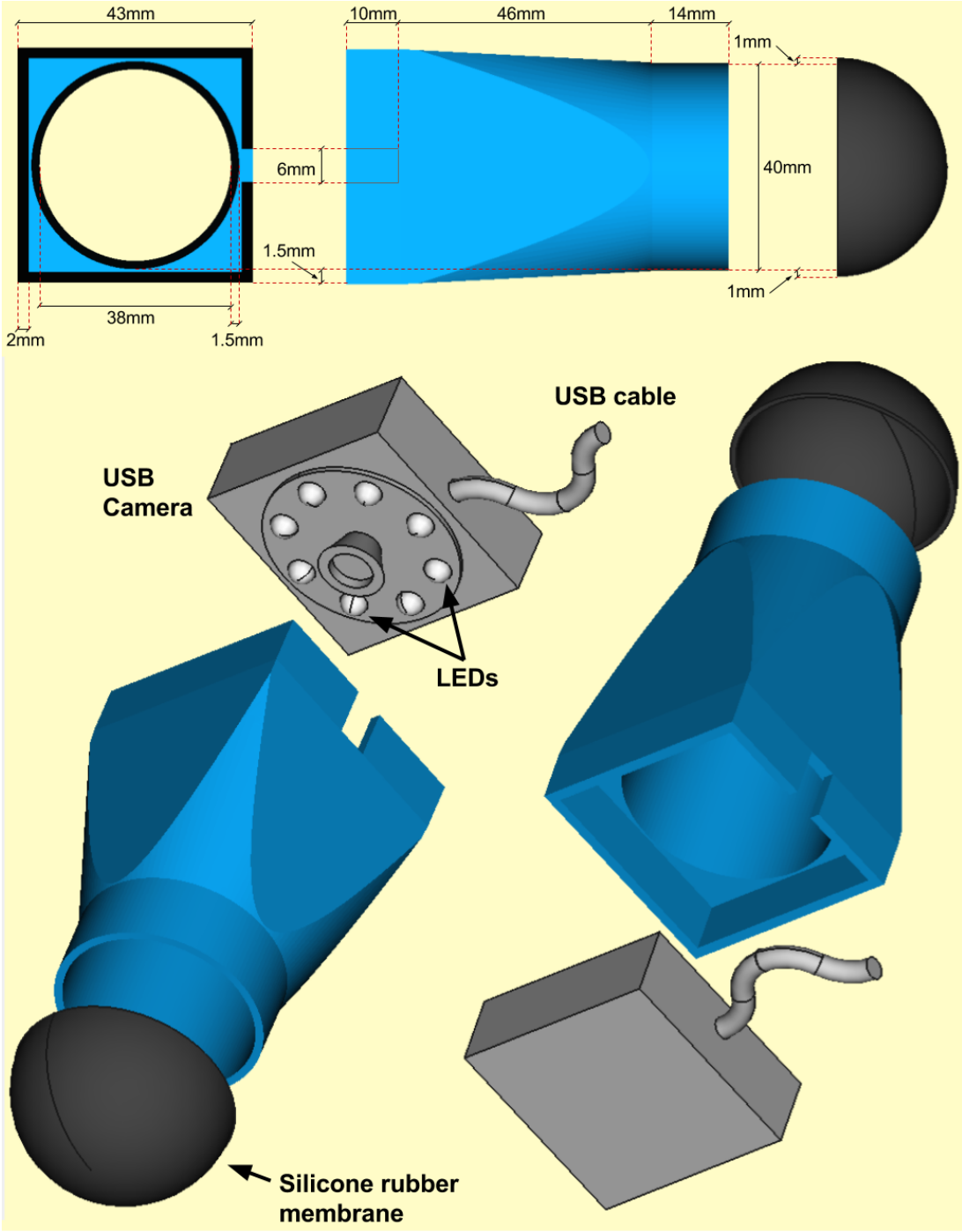


Figure 3-1: *Schematics of the BathTip's dimensions (top) and isometric view of the sensor assembly (bottom). See Section 3.2 for details.*

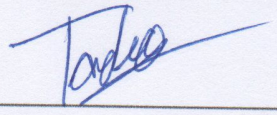
that the raw image is used as an input to the various linear encodings. The high contrast between the papillae (white) and the background (black) in the TacTip mean that small perturbations to the surface may result in large perturbations to the raw image vector. It would perhaps make for a better encoding to use tracking on the papillae and to use their location and displacement as input (such as in [72]), instead of the raw image. Therefore, this result should be considered as a way of validating the BathTip, not as a way of stating a superior capability.

3.4 Errata

The paper citation number referring to the work of Barron-Gonzalez and Prescott (in-paper reference [2]), incorrectly states that their work was published in ICRA 2013. Their work was in fact published in TAROS 2013, the correct reference is found in the bibliography of this thesis [6].

3.5 Paper: Tactile Features: Recognising touch sensations with a novel and inexpensive sensor

The sensor design, data encoding, and shape recognition experiment and results were published [24] as a paper at TAROS (Towards Autonomous Robotics Systems Conference), achieving the 'Best Student Paper Prize'. The Statement of Authorship Form and the paper can be found next.

This declaration concerns the article entitled:									
Tactile Features: Recognising touch sensations with a novel and inexpensive tactile sensor									
Publication status (tick one)									
draft manuscript		**Submitted**		**In review**		**Accepted**		**Published**	✓
Publication details (reference)	T. Corradi, P. Hall, and P. Iravani, Tactile features: Recognising touch sensations with a novel and inexpensive tactile sensor, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8717 LNAI, Springer Verlag, 2014, pp. 163-172								
Candidate's contribution to the paper (detailed, and also given as a percentage).	The candidate contributed to/ considerably contributed to/predominantly executed the… Formulation of ideas: 90%. I designed the sensor, had the idea of shape recognition, and the ideas on how to test the sensor and encodings. My supervisor proposed one of the possible encodings to use. Design of methodology: 100%. I designed the experiments, the measures of success, the sensor, and the entire process. Experimental work: 100%. I conducted all experiments, evaluations and comparisons. Presentation of data in journal format: 90%. I collated data, wrote all drafts, and submitted the work. My supervisor provided feedback on drafts.								
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.								
Signed							**Date**	03/04/2018	

Tactile features: recognising touch sensations with a novel and inexpensive tactile sensor

Tadeo Corradi, Peter Hall, and Pejman Iravani

University of Bath, Bath, UK
t.m.corradi@bath.ac.uk

Abstract. A simple and cost effective new tactile sensor is presented, based on a camera capturing images of the shading of a deformable rubber membrane. In Computer Vision, the issue of information encoding and classification is well studied. In this paper we explore different ways of encoding tactile images, including: Hu moments, Zernike Moments, Principal Component Analysis (PCA), Zernike PCA, and vectorized scaling. These encodings are tested by performing tactile shape recognition using a number of supervised approaches (Nearest Neighbor, Artificial Neural Networks, Support Vector Machines, Naive Bayes). In conclusion: the most effective way of representing tactile information is achieved by combining Zernike Moments and PCA, and the most accurate classifier is Nearest Neighbor, with which the system achieves a high degree (96.4%) of accuracy at recognising seven basic shapes.

Keywords: Haptic recognition, tactile features, tactile sensors, supervised learning

1 Introduction

The aim of this paper is to find an accurate low-dimensional representation of a tactile image perceived by a novel tactile sensor developed by us, these representations are from now on referred to as 'encodings'. Tactile sensors and tactile information encoding have been focus of much research lately [5]. Whilst numerous standards exist in Computer Vision, there is no consensus on the best approach to encoding tactile sensing information [5], and the only tactile database known to us [24] is limited to a single sensor type. Unlike visual information, haptic information can be distributed over a potentially unknown geometry [5] (for example a single robotic hand can be fitted with many different combinations of tactile sensors), so the equivalent problem to 'camera calibration' is a significantly more difficult task. Whilst the majority of efforts have gone to low resolution sensor pads [2], [16], [19], [20], [26], a new biologically inspired sensor design, called the TacTip [3] aims to provide higher resolution whilst remaining inexpensive. This paper presents a similar, simplified, low cost tactile sensor and evaluate its accuracy recognising 7 basic tactile shapes (Corner, Cylinder, Edge, Flat-to-Edge, Flat, Nothing, Point), comparing a selection of encodings and a range of supervised classifiers.

2 Related Work

Tactile sensors can be designed using a variety of techniques, perhaps the most popular being resistive sensors [28]; but also including magnetic, piezo-electric, capacitive and others [5]. A large amount of effort has been put into texture recognition [7], [11], [14], [25], since texture is usually difficult to capture from vision alone. The most direct approach to tactile feature classification is to use the tactile images with no encoding and use a simple distance metric [23]. Recently, there have been several projects involving recognition by grasping using Pattern Recognition techniques to find the best dimensionality reduction function for tactile information. Early approaches focused on tailored designs [1] or classical Artificial Neural Networks (ANNs) [27]. More recently, PCA, moment analysis and binary (contact/no contact) have been compared in a system that integrates tactile and kinesthetic information for object recognition [8], finding that the use of central moments outperforms other encodings. A variation on Self-Organizing Maps (SOMs) [13] has been developed and applied to fusing propioceptive and tactile input for object recognition [12]. PCA and SOMs have been used to extract tactile features which were then used for object recognition [16]. Novel recursive gaussian kernels have been used to encode the various stages of contact during grasping leading to a robust online classifier [26].

2.1 The TacTip

Most previous studies are based on pressure sensor arrays. An innovative biologically inspired sensor was proposed recently [3] which uses a flexible hemispherical membrane with internal papillae which move as the membrane deforms whenever it touches an object. A digital camera records and transmits the image of the displaced papillae (see right side of Fig. 1). This sensor, called the TacTip, was shown to achieve a high degree of accuracy in sensing edges [4] to a point where a small object is clearly identifiable by a human from its tactile image and has been theoretically shown to have potential in tele-surgery [21]. More recently it has also been successfully used to identify textures [29]. The new sensor presented by this paper is an adaptation of the TacTip. No papillae nor internal gel is needed (significantly simplifying the sensors manufacture process and cost) and the shading pattern of light is used as input, instead of the papillae locations. This paper shows that the new sensor is effective at recognising tactile shapes.

3 Sensor Specification

3.1 Design

The new sensor consists of an opaque silicone rubber hemispherical membrane of radius 40mm and thickness 1mm, mounted at the end of a rigid opaque cylindrical ABS tube. At the base of the tube, there is a PC web-cam equipped with 8 white LEDs. The LEDs illuminate the rubber, the shading pattern of the image changes as the rubber makes contact with various surfaces (see Fig. 1).

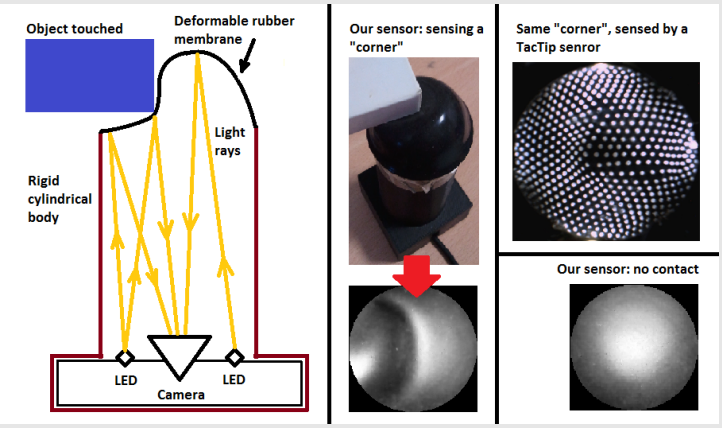


Fig. 1. The new tactile sensor design (left). The main body is 3D printed in ABS. The tip is a 1mm thick silicone rubber hemisphere. At the base (not visible) there is a USB eSecure©web-cam with 8 LEDs illuminating the inside of the silicone hemisphere (bottom right). As the tip makes contact with an object, it deforms resulting in a specific shading pattern (middle). As a comparison, the same tactile shape as perceived by a TacTip is shown (top right).

4 Methods

4.1 Preprocessing: Discrete Derivative

The shading pattern is related to the angle between the membrane's normal and the light rays going to the camera. Therefore drastic changes in luminosity are to be expected whenever the discrete spatial derivative of the normal of the surface is highest, that is where the rubber is most sharply bent (see Fig. 2). This concept motivates the analysis of the images' discrete derivative's magnitude matrix $D(I)$, defined, for any square image matrix $I \in \mathbb{R}^{w \times w}$, as:

$$D(I)_{i,j} := +\sqrt{(I_{i-1,j} - I_{i+1,j})^2 + (I_{i,j-1} + I_{i,j+1})^2}, \quad \forall i,j \in [1, w-1] \qquad (1)$$

In the experiments described below, encodings will be applied to the raw image received by the camera, and to the magnitude of its discrete derivative, $D(I)$.

4.2 Rotationally Invariant Encodings

Due to the circular geometry of the sensor image, a rotation invariant encoding was required. Five alternatives were explored: Hu moments [10], Zernike Moments [30], Principal Component Analysis (with regularized rotation), Zernike-PCA (PCA applied to the Zernike moments), and image scaling.

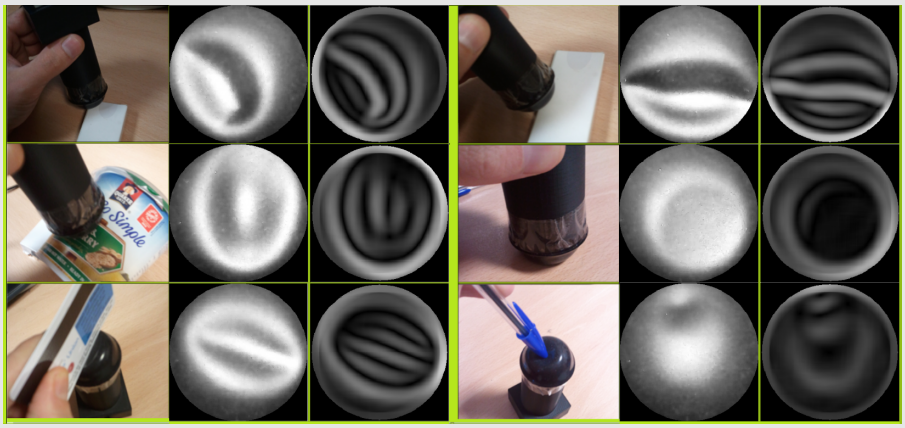


Fig. 2. Examples of occurrences of 6 of the 7 basic tactile shapes (the 7th is "nothing", in Fig. 1) (left columns), and their corresponding shading pattern (middle columns) and the magnitude of its first spatial derivative $D(I)$ (right columns). From the top left, downwards: Corner, Cylinder, Edge, Flat-to-Edge, Flat, Point.

Hu Moments Hu moments are special combination of central moments which aim to be invariant to rotation, translation and scale (for details see [10]). The implementation used here was the one by [15], who have demonstrated the use of Hu moments in effective feature extraction on edge images for object recognition.

Zernike Moments A Zernike Moment is the element-wise product of an image with a Zernike polynomial evaluated at the locations of the pixels of the image, rescaled to circumscribe a unit disk.

Definition 1. *Let $m \geq n$ be non-negative integers, and let $0 \leq \phi \leq 2\pi, 0 \leq \rho \leq 1$ define a polar coordinate system. Then the even and odd Zernike polynomials are defined as:*

$$Z_n^m(\rho, \varphi) = R_n^m(\rho)\, \cos(m\varphi) \tag{2}$$

$$Z_n^{-m}(\rho, \varphi) = R_n^m(\rho)\, \sin(m\varphi), \tag{3}$$

Which can be indexed by:

$$Z_j = Z_{n(j)}^{m(j)} \tag{4}$$

Where $m(j), n(j)$ are Noll's indices (See Table 1) of Zernike polynomials [17], and

$$R_n^m(\rho) = \sum_{k=0}^{(n-m)/2} \frac{(-1)^k\,(n-k)!}{k!\,((n+m)/2-k)!\,((n-m)/2-k)!}\,\rho^{n-2\,k} \tag{5}$$

Table 1. First ten Noll indeces [17] to compose a linear sequence of Zernike polynomials.

j	1	2	3	4	5	6	7	8	9	10
n(j)	0	1	1	2	2	2	3	3	3	3
m(j)	0	1	-1	0	-2	2	-1	1	-3	3

Now, the d^{th} Zernike Moment of an image M is given by:

$$Zer_d(M) = \left| \sum_{i,j \in \{i^2+j^2 \leq n^2/2\}} M(i,j) Z'_d(i,j) \right| \tag{6}$$

Where,

$$Z'_d(i,j) := Z_j \left(\frac{\sqrt{(i^2+j^2)}}{\frac{\sqrt{2}}{2}n}, arctan\left(\frac{j-n/2}{i-n/2}\right) \right) \tag{7}$$

PCA and Zernike-PCA In the third encoding, the orientation of each image was computed (from central moments) and the image was rotated so as to regularize its orientation. Then PCA was performed on the vectorised images. The fourth encoding, Zernike-PCA, was simply applying PCA to the Zernike Moments of all images. In both of these, the dimensionality reduction matrix was computed on training data and used for both the training dataset and the testing dataset.

Scaling (Vectorized) For the fifth encoding, image orientations are regularised first, then images are resized by averaging pixel intensities, into a much smaller resolution (up to 13 by 13 pixels, from an original resolution of 300 by 300). The resulting images are vectorized, so for example, a 13-by-13 image, is converted into a 1-by-13^2 vector, by concatenating the pixel columns.

4.3 Encoding Evaluation

Each of these encodings was applied to a training dataset of 175 images, labelled from 1 to 7, corresponding to the tactile shapes they represented (see Fig. 2). Each encoding will produce a different set of data clusters. Good encodings will result in clusters which are spatially conglomerate: vectors corresponding to images of equal label will be close together and those with different labels will be far apart. One way of measuring this property is the Davies-Bouldin index in L^2 [6], defined below. Lower values of this index represent more distinctive clusters.

Definition 2. *Let $d(a, b)$ represent the euclidean distance metric. Let X be a set of vectors of dimension d, partitioned into k disjoint clusters, $X = \bigcup_{i=1}^{k} X_i$. Let c_i be the centroid of cluster X_i. The Davies−Bouldin index is given by:*

Gives lkinear indexing for zernike polynomials [17]

$$D = \frac{1}{k} \sum_{i=1}^{k} \max_{j:i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i - c_j)} \right) \tag{8}$$

Where,

$$\sigma_i := \sqrt{\frac{1}{|X_i|} \sum_{x \in X_i} d(x - c_i)^2} \tag{9}$$

Classifiers Cross Validation As a second way of judging the suitability of a particular encoding is to train a supervised classifier given the known labels and to test their accuracy at predicting the labels of the encoded data. The measure used here is the 5-fold cross validation accuracy, defined as the average percentage of correct classifications performed by a given classifier trained with $\frac{4}{5}$ of the labelled data and tested on the remaining $\frac{1}{5}$ of the data. The process is repeated 5 times so that all data is used for testing. This method was applied to the following classifiers:

- Nearest Neighbor classifier
- Artificial Neural Network with a single 7 neuron hidden layer, trained using backpropagation.
- A group of seven binary Support Vector Machines (one per label) used in conjunction, arbitrarily choosing the largest label id, if more than one returned a positive classification.
- A simple Naive Bayes classifer, using Kernel Density Estimation (KDE) [18], [22].

For the implementation of these four algorithms, and for the simulations described in this paper, MATLAB[1] was used.

5 Results

Seven basic tactile shapes were defined: Corner, Cylinder, Edge, Flat-to-Edge, Flat, Nothing, and Point. Using the new sensor images were manually captured, resulting in 70 sample frames of each one (see Fig. 2). Data was split: 175 images were used for training (selecting the optimum encoding vector size), and the remaining 175 images for validation. Each one of the encodings defined in Section 4.2 was applied to each training image and the magnitude of its discrete derivatives (as described in Section 4.1). Then two tests were performed: cluster evaluation and classifier evaluation.

[1] MATLAB©, Statistics Toolbox and Neural Network Toolbox Release 2013b, The MathWorks, Inc., Natick, Massachusetts, United States.

5.1 Cluster evaluation

First, the Davies-Bouldin index was computed on the training data data (175 images) to find the optimum number of components to use in each encoding (number of principal components, number of zernike polynomials, etc.). This parameter (number of components) is then fixed and the Davies-Bouldin index is computed on the remaining 175 images (the validation dataset). Table 2 shows the result. Zernike moments combined with PCA seem to produce the most distinct clusters under this criteria. Cluster formation using the new sensor seems superior with respect to the TacTip using this measure. This is possibly due to the fact that papillae displacements mean that small perturbations in the object surface translate into significant non-linear changes in the image.

Table 2. Davies−Bouldin index (described in Section 4.3) computed for the clusters resulting from the different encodings. They represent the distinctiveness of a cluster, smaller numbers represent better defined clusters.

Encoding	Applied to Image (Our sensor)	Applied to Image (TacTip sensor)	Applied to D(Image) (Our sensor)	Applied to D(Image) (TacTip sensor)
Hu Moments	5.3	10.4	5.1	13.2
Zernike M.	2.0	2.5	1.9	3.8
PCA	2.6	5.9	1.8	5.4
ZernikePCA	1.5	2.6	1.4	2.9
Scale (Vect.)	37.1	37.9	10.4	1378.8

5.2 Classifier evaluation

Each one of the classifiers described in Section 4.3 is now trained. Using 20 iterations of randomized 5-fold cross validation on the training dataset the optimal vector sizes for each encoding and classifier are obtained. Then, the process is repeated on the validation dataset, but only using these optimum vector sizes. Figure 3 shows the accuracy of each encoding/classifier pair.

Zernike PCA applied directly to the image outperforms other encodings in general. In terms of classifiers, Nearest Neighbor is the overall best for both sensors, reaching an accuracy on the validation dataset of 96.4%. It must be born in mind that Nearest Neighbor classifiers using cross validation are prone to data twinning (bias if similar data are present in a dataset). To reduce the effects, a small value for k (5) was used in k-fold cross validation, together with randomisation and multiple trials; furthermore, separate dataset were used for training and validation. Nevertheless, if data twinning is likely to be an issue in further applications, it may be advisable to use Naive Bayes (KDE).

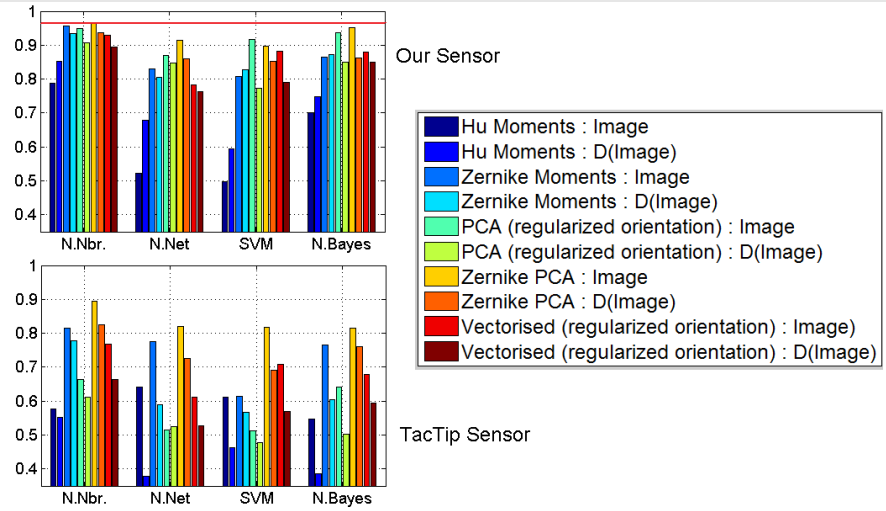


Fig. 3. Randomized 5-fold cross validation accuracy for the 7 basic tactile shapes (higher is better, 1 is 100% perfect recognition). Input set of 175 labelled tactile images, corresponding to 7 clusters. Comparison between our sensor and the TacTip, using four different encodings as classified by four different supervised algorithms.

There is no significant difference between the performance of any encoding/classifier pairing when comparing their use on the image and on its derivative. This may be due to the fact that the discrete derivative only loses base intensity information, which is a single degree of freedom over images which are 90000-dimensional. The accuracy achieved with our sensor is slightly higher to the one with the TacTip, for these particular choices of encodings and classifiers. Once again, the non-linearity introduced by papillae is potentially a factor, and so the comparison is by no means exhaustive in scope.

6 Conclusions

This paper presented a novel, simple and inexpensive tactile sensor based on shading resulting from the deformation of a rubber membrane. Various encodings were tested on the input images and on their discrete derivatives. For each encoding, the accuracy of a selection of classifiers was tested, by performing tactile shape recognition. The new sensor is capable of distinguishing between these shapes, the most accurate encoding is Zernike Moments combined with PCA, applied directly to the input image. The most accurate classifier is Nearest Neighbor, which reaches a classification accuracy of 96.4%. Our sensor performed slightly better than the TacTip in these tests, which is remarkable considering the simplicity of our sensor's design. However it must be stressed that other approaches may very well favor the TacTip. The discrete localization of the papillae may be a disadvantage in linear encodings, but it can be an advantage in general, as it is more resilient to image noise and less dependent on calibration of camera parameters. Only pattern recognition was discussed in this paper, it may be of interest to use "shape from shading" [9] to reconstruct the exact shape of the deformed hemisphere. Further work should also focus on this sensor's potential for object recognition.

Acknowledgments. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK. We would like to thank Bristol Robotics Lab[2] for lending us the TacTip sensor.

References

1. Allen, P.K.: Integrating vision and touch for object recognition tasks. The International Journal of Robotics Research 7(6), 15–33 (Jan 1988)
2. Barron-Gonzalez, H., Prescott, T.: Discrimination of social tactile gestures using biomimetic skin. In: IEEE International Conference on Robotics and Automation. Karlsruhe, Germany (2013)
3. Chorley, C., Melhuish, C., Pipe, T., Rossiter, J.: Development of a tactile sensor based on biologically inspired edge encoding. In: International Conference on Advanced Robotics, 2009. ICAR 2009. pp. 1–6 (2009)

[2] www.brl.ac.uk

4. Chorley, C., Melhuish, C., Pipe, T., Rossiter, J.: Tactile edge detection. In: 2010 IEEE Sensors. pp. 2593–2598 (2010)
5. Dahiya, R., Mittendorfer, P., Valle, M., Cheng, G., Lumelsky, V.: Directions toward effective utilization of tactile skin: A review. IEEE Sensors Journal 13(11), 4121–4138 (2013)
6. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1(2), 224–227 (Apr 1979)
7. Decherchi, S., Gastaldo, P., Dahiya, R., Valle, M., Zunino, R.: Tactile-data classification of contact materials using computational intelligence. IEEE Transactions on Robotics 27(3), 635–639 (2011)
8. Gorges, N., Navarro, S., Goger, D., Worn, H.: Haptic object recognition using passive joints and haptic key features. In: 2010 IEEE International Conference on Robotics and Automation (ICRA). pp. 2349–2355 (2010)
9. Horn, B.K.P., Brooks, M.J. (eds.): Shape from Shading. MIT Press, Cambridge, MA, USA (1989)
10. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory 8(2), 179–187 (Feb 1962)
11. Jamali, N., Sammut, C.: Majority voting: Material classification by tactile sensing using surface texture. IEEE Transactions on Robotics 27(3), 508–521 (2011)
12. Johnsson, M., Balkenius, C.: Sense of touch in robots with self-organizing maps. IEEE Transactions on Robotics 27(3), 498–507 (2011)
13. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1), 59–69 (Jan 1982)
14. Liu, H., Song, X., Bimbo, J., Seneviratne, L., Althoefer, K.: Surface material recognition through haptic exploration using an intelligent contact sensing finger. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 52–57 (2012)
15. Mercimek, M., Gulez, K., Mumcu, T.V.: Real object recognition using moment invariants. Sadhana 30(6), 765–775 (Dec 2005)
16. Navarro, S., Gorges, N., Worn, H., Schill, J., Asfour, T., Dillmann, R.: Haptic object recognition for multi-fingered robot hands. In: 2012 IEEE Haptics Symposium (HAPTICS). pp. 497–502 (2012)
17. Noll, R.J.: Zernike polynomials and atmospheric turbulence. Journal of the Optical Society of America 66(3), 207–211 (Mar 1976)
18. Parzen, E.: On estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076 (1962)
19. Pezzementi, Z., Plaku, E., Reyda, C., Hager, G.: Tactile-object recognition from appearance information. IEEE Transactions on Robotics 27(3), 473–487 (2011)
20. Ratnasingam, S., McGinnity, T.: A comparison of encoding schemes for haptic object recognition using a biologically plausible spiking neural network. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3446–3453 (2011)
21. Roke, C., Melhuish, C., Pipe, T., Drury, D., Chorley, C.: Deformation-based tactile feedback using a biologically-inspired sensor and a modified display. In: Gro, R., Alboul, L., Melhuish, C., Witkowski, M., Prescott, T.J., Penders, J. (eds.) Towards Autonomous Robotic Systems, pp. 114–124. No. 6856 in Lecture Notes in Computer Science, Springer Berlin Heidelberg (Jan 2011)
22. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics 27(3), 832–837 (Sep 1956)

23. Schneider, A., Sturm, J., Stachniss, C., Reisert, M., Burkhardt, H., Burgard, W.: Object identification with tactile sensors using bag-of-features. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009. pp. 243–248 (2009)
24. Schopfer, M., Ritter, H., Heidemann, G.: Acquisition and application of a tactile database. In: 2007 IEEE International Conference on Robotics and Automation. pp. 1517–1522 (2007)
25. Sinapov, J., Sukhoy, V., Sahai, R., Stoytchev, A.: Vibrotactile recognition and categorization of surfaces by a humanoid robot. IEEE Transactions on Robotics 27(3), 488–497 (2011)
26. Soh, H., Su, Y., Demiris, Y.: Online spatio-temporal gaussian process experts with application to tactile classification. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4489–4496 (2012)
27. Taddeucci, D., Laschi, C., Lazzarini, R., Magni, R., Dario, P., Starita, A.: An approach to integrated tactile perception. In: 1997 IEEE International Conference on Robotics and Automation. vol. 4, pp. 3100–3105 vol.4 (1997)
28. Weiss, K., Worn, H.: The working principle of resistive tactile sensor cells. In: Mechatronics and Automation, 2005 IEEE International Conference. vol. 1, pp. 471–476 Vol. 1 (2005)
29. Winstone, B., Griffiths, G., Pipe, T., Melhuish, C., Rossiter, J.: TACTIP - tactile fingertip device, texture analysis through optical tracking of skin features. In: Lepora, N.F., Mura, A., Krapp, H.G., Verschure, P.F.M.J., Prescott, T.J. (eds.) Biomimetic and Biohybrid Systems, pp. 323–334. No. 8064 in Lecture Notes in Computer Science, Springer Berlin Heidelberg (Jan 2013)
30. Zernike, V.: Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. Physica 1(7-12), 689–704 (May 1934)

Chapter 4

Tactile object recognition

4.1 Motivation: from tactile shapes to tactile object recognition

Chapter 3 corroborated that the simplified sensor was apt for shape recognition and identified a way to encode data using Zernike moments and PCA. The next step was to create an algorithm on top of this features, for object recognition. A single robotic hand can be fitted with many different combinations of tactile sensors and each such configuration would be unique. It is therefore not surprising that tactile databases are a relatively new phenomenon [106, 7, 8, 31]. These databases are all limited to pressure sensor arrays, and all except the first are based on some form of grasping. Therefore a new data set would be needed. Following the nature and number of objects used in similar work (e.g. [91, 105]), the type of objects chosen was household objects (such as bottles, books, etc.) and the sample size was set to 10. This would give an early indication of whether object recognition (and perhaps classification) would be feasible.

4.2 Summary: data collection and tactile object representation

The aim of this stage of the project was to design object representations which are able to recognise simple household objects. A procedure for tactile exploration of 10 such objects was designed, using the BathTip sensor, mounted on a robotic arm. The arm would sequentially move inwards towards the object from

a randomised approach angle until contact was detected, recording the resulting tactile image. Detection of contact was performed using an icub FTSENS 6-axis force-torque sensor[1], mounted between the robotic arm end-effector and the tactile sensor. When a compressing (z-axis) force of at least 0.74N was read, the arm would stop and a tactile image would be recorded. The value of 0.74N was chosen after manual experiments to discern a suitably large force so the sensor's rubber membrane would be deformed sufficiently but not so large as to endanger the sensor's integrity. Only small parts of an object can be sensed at a time. Therefore, several such tactile images must be considered simultaneously. The relative location and orientation of the contact position are not straightforward to compute, and prone to large relative errors, considering the object's pose may be perturbed by the tactile interaction.

As a consequence of these considerations and the nature of the data, a bag-of-words model was adopted (similar to [105]), discarding information about the position and orientation of the sensor. Building on the results of the first publication (Chapter 3), Zernike-PCA encoding was used for tactile information. In order to be able to represent objects from tactile perceptions, a novel likelihood function was devised, which models the probability of each Zernike-PCA vector, given an object's label. Thereafter, object recognition was performed by maximising the marginal likelihood of test data. The likelihood function designed is a normalised sum of Gaussian probability densities, with means equal to the training samples vectors, and covariance equal to the training set's covariance matrix. This is similar to a Gaussian Mixture Model (GMM) [82] with the number of components equal to the number of training vectors, and component coefficients all equal to one. It is also similar to Multi-Variate Kernel Density Estimation (MVKDE) [13] with a normal kernel, and bandwidth set to the covariance of the data. Both GMM and MVKDE were tested on validation data sets and classification results were significantly poorer than the model here proposed.

The inference process is described as Bayesian, since, when a test object is classified, there are multiple tactile readings, and to obtain a probability distribution over object labels, the following equation is used.

$$P(C|Y) = \alpha \prod_{j=1}^{m} P(Y_j|C)P(C) \tag{4.1}$$

[1] http://www.icub.org/

Where m is the number of tactile touches considered jointly, Y_j is the j^{th} tactile vector (the Zernike-PCA vector resulting from a tactile image), C is an object class, and α is a normalising constant. This can also be framed in terms of Bayesian updates, where there is an initial prior probability for a class, $P_0(C)$, and m updates steps are performed, one for each tactile vector read. Each step uses Bayes' rule to update the believed probability, given the tactile input.

$$P_j(C) := P(C|Y_j) = \frac{P(Y_j|C)P_{j-1}(C)}{P(Y_j)},\ j = 1, ..., m \tag{4.2}$$

After m updates, the final posterior probability for class C, $P_m(C)$ is given by:

$$P_m(C) = P_0(C) \prod_{j=1}^{m} \frac{P(Y_j|C)}{P(Y_j)} \tag{4.3}$$

Which is equivalent to Equation (4.1), since the denominator is constant, i.e. $\alpha = \prod_{j=1}^{m} P(Y_j)^{-1}$.

4.3 Results: state-of-the-art non-grasping tactile recognition

Tactile recognition within the 10 object data set ranged from 0.5 to 0.95 depending on the number of touches considered at test time. At the time, this was the highest accuracy reported in comparable (tactile only non-grasping) experiments. There were indications that the approach could be used for classification of unseen objects: 4 new objects were correctly classified, but further work was needed at this stage to corroborate that hypothesis.

One remarkable result was the presence of high uncertainty when the system aimed to classify previously this unseen object. Furthermore, in the other four cases, the correct object label obtained a high value posterior probability with few touches. This points to a potential further approach using Sequential Analysis [122], where the test may be stopped early, if sufficient evidence is considered to be already gathered to make a decision as to the identity of the object being recognised.

4.4 Paper: Bayesian tactile object recognition: learning and recognising objects using a new inexpensive tactile sensor

The details of the data collection, model definition, experiment and results were published at the International Conference on Robotics and Automation (ICRA) [25]). The Statement of Authorship Form and the paper can be found next.

This declaration concerns the article entitled:									
Bayesian tactile object recognition: Learning and recognising objects using a new inexpensive tactile sensor									
Publication status (tick one)									
draft manuscript		**Submitted**		**In review**		**Accepted**		**Published**	✓
Publication details (reference)	T. Corradi, P. Hall and P. Iravani, "Bayesian tactile object recognition: Learning and recognising objects using a new inexpensive tactile sensor, in 2015 IEEE International Conference on Robotics and Automation (ICRA), vol. 2015-June, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 3909-3914.								
Candidate's contribution to the paper (detailed, and also given as a percentage).	The candidate contributed to/ considerably contributed to/predominantly executed the… Formulation of ideas: 100%. I proposed the experiment, the hypothesis, and the technological and methodological approach. Design of methodology: 90%. I decided what data to collect and how, and how. I designed the algorithms that would be used to recognise objects. My supervisors provided ideas regarding the force-torque sensor and critical checking of the likelihood model. Experimental work: 95%. I prepared and carried out all data collecting and programming. My supervisors provided support in programming the control for the robotic arm and provided some initial code for zernike polynomial calculation. Presentation of data in journal format: 85%. I wrote all drafts, created all figures, and submitted the paper. I attended and presented at the conference. My supervisors gave feedback on various drafts, including substantial rephrasing of abstract and introduction.								
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.								
Signed	Tadeo					**Date**	03/02/2018		

Bayesian Tactile Object Recognition: learning and recognising objects using a new inexpensive tactile sensor*

Tadeo Corradi[1], Peter Hall[2] and Pejman Iravani[1]

***Abstract*— We present a Bayesian approach to tactile object recognition that improves on state-of-the-art in using single-touch events in two ways. First by improving recognition accuracy from about 90% to about 95%, using about half the number of touches. Second by reducing the number of touches needed for training from about 200 to about 60. In addition, we use a new tactile sensor that is less than one tenth of the cost of widely available sensors. The paper describes the sensor, the likelihood function used with the Naive Bayes classifier, and experiments on a set of ten real objects. We also provide preliminary results to test our approach for its ability to generalise to previously unencountered objects.**

I. INTRODUCTION

A Bayesian system for object learning and recognition using purely tactile, orientation independent information is presented. A novel, inexpensive sensor is used, mounted on a robotic arm which learns in an automatic manner to recognise objects outperforming state-of-the-art. We also provide some evidence that the system can recognise previously unseen objects.

The system learns and recognises objects from single-touch events using a newly developed sensor [1]. Tactile sensations are encoded using Zernike Moments and objects are modeled by a sum of Gaussian distributions. The approach presented does not use the orientation information of the objects and requires only a very limited number of training samples, making a substantial improvement over previous work. A fully automated robot system (depicted in Fig. 1) was constructed to learn the tactile appearance of 10 household objects and to recognise these with an accuracy of 87% after 15 touches and 95% after 30 touches.

II. RELATED WORK

A. Tactile sensors

Tactile sensors can be designed using a variety of techniques, the most common being piezo-resisitive sensors, conductive polymers, or capacitive sensors [2]. The most widely used in robotics include the impedance based BioTac [3], the Weiss tactile array [4], and the capacitive array based DigiTact [5], all of which have a price tag exceeding USD 700. Recently, efforts have been made at creating cheaper and more accessible sensors. The TakkTile TakkArray [6] is an open source and open hardware sensor based on an array of MEMS barometers, it has a retail price of USD 500, and their material cost is approximately USD 200. The TacTip [7] aims to provide higher resolution whilst remaining inexpensive as they can be non-professionally manufactured (material cost is approximately USD 200). It is a biologically inspired tactile sensor based on the deformation of a silicone rubber hemispherical surface and the consequential displacement of a number of internal papillae. A digital camera is used to observe this displacement.

Fig. 1. The new tactile sensor mounted on a KUKA KR5-sixx-R650 robotic arm, currently exploring the tactile appearance of a mug.

B. Recognition by grasping

Recently, there have been several projects involving recognition by grasping using machine learning techniques. Principal Component Analysis (PCA), Self Organizing Maps and Artificial Neural Networks have been combined to process the output of Weiss tactile sensory arrays attached to a number of robotic end-effectors, to recognize household objects [8]. Novel recursive Gaussian kernels have been designed to encode the various stages of contact during grasping leading to a robust on-line system capable of learning new models and classifying objects in real time [9]. The most accurate system, to the best of our knowledge, is the one developed by [10]. They extends HMP (Hierarchical Matching Pursuit, a multi-layer hierarchical feature learning system) to include temporal information. They test their method on 6 tactile databases and produce an accuracy of between 80% and 100%. Whilst it is evident that combining proprioceptive with tactile information is likely to yield better results than either modality alone [11], [12], using grasp limits the size

*This work was funded by the Engineering and Physical Sciences Council (EPSRC), UK

[1]Department of Mechanical Engineering, University of Bath, Claverton Down, Bath, BA27AY, UK t.m.corradi@bath.ac.uk

[2]Department of Computer Science, University of Bath, Claverton Down, Bath, BA27AY, UK

of the object to be identified, requires a robotic hand, and requires a grasp to be achieved.

C. Single contact tactile recognition

Recognition using a single touch at a time is a possible solution which remains relatively unexplored. As far as we know, the best results so far are achieved by [13], requiring 60 touches to converge to 90% recognition accuracy, using 200 touches for training, over a set of 5 objects.

The most common approaches for single contact tactile object recognition are voxel based or point clouds [14], [15], [16]. Recently, a very efficient and accurate combination of both was developed [17], which is able to model the object shape and the uncertainty about occupied space. They achieve above 80% accuracy in recognition over a set of 45 objects, and from only 10 touches; however, object 3D models are required in advance. Voxel representations and point-clouds provide a natural way of representing tactile information about objects, but they can be cumbersome in terms of computational power for recognition, as they usually comprise a large number of points/voxels whose matching to a database can be complex, and are prone to noise which is difficult to model. Attempts to address these problems include merging points that are close into a probability point modelled by a Kalman filter [18], and clustering to subdivide the point cloud into regions which are then encoded as features [19].

D. Appearance based tactile-only recognition

One of the first attempts at a tactile-only recognition is [20], which uses geometric features such as lines and points and their evolution over time. Their accuracy recognising objects is high (83%), however the number of shapes is only 6 and they are very basic predefined geometric solids (cylinder, cone, etc.). The two notable recent pieces of research which most closely relate to our study are the work of Schneider et al. [21], and the work of Pezzementi et al. [13].

The first [21], involves the repeated application of a two fingered grasps using a gripper equipped with Weiss tactile array sensors. Features are extracted, then a bag-of-features approach is used to recognise household and industrial objects. They use an information theoretic approach for maximum expected information gain to inform grasping position. They obtain an accuracy of 84.6% in recognition, using 830 tactile images for training and 16 to 20 tactile images in the testing set. The object pose is strictly known and fixed (small translation variance is tolerated). It could be argued that this work uses proprioception (they know the height of the gripper) and thus is not purely appearance based.

Pezzementi et al. [13] use simulations to compare various methods of feature extraction, and create clusters of these features to compile feature histograms to be compared for object recognition. Most of their testing is performed in simulation using 3D models of objects. The physical testing was done using DigitTact sensors over a set of 5 objects (the context was recognition of plastic letters) using a predefined exploring routine. They use 200 samples for training and 100 for testing. The accuracy in these physical experiments reaches 90% for one of their feature choices after approximately 60 touches. It would be interesting to see this system tested on a larger set of objects, since its simulated performance is quite good.

III. Sensor and Tactile Data Representation

The new sensor [1] used in this paper is based on the same principle as the TacTip. However, it has neither papillae nor internal gel. Instead it has a plain black smooth opaque silicone rubber hemispherical membrane of radius 40mm and thickness 1mm, mounted at the end of a rigid opaque encasing for the digital camera, 3D printed in ABS [1]. The camera has a resolution of 640 by 480 pixels, and incorporates a set of 8 white LEDs. The shading pattern of light is used as input. When the sensor is in contact with an object, the shading pattern on the membrane changes accordingly (see Fig. 2). In recent work, it was shown to recognise seven basic shapes with over 95% accuracy [1].

Due to the circular geometry of the sensor image, a rotationally invariant representation was required. In previous work, a number of encoding methods were compared and it was suggested that Zernike Moments together with PCA achieved the best performance [1]. Zernike Moments have been shown to be useful when scale, rotation and translation invariances are sought [22], and have been successfully used for basic shape recognition [23]. Zernike moments here refers to the absolute value of the inner product of a vectorised image with a vectorised Zernike polynomial, a set of radial complex polynomials defined on the unit disk (see Fig. 3).

Let $m \geq n$ be non-negative integers, and let $0 \leq \phi \leq 2\pi, 0 \leq \rho \leq 1$ define a polar coordinate system. Then the

[1]3D model of the tactile sensor encasing, and links to the other components are available at: `https://github.com/Exhor/bathtip`

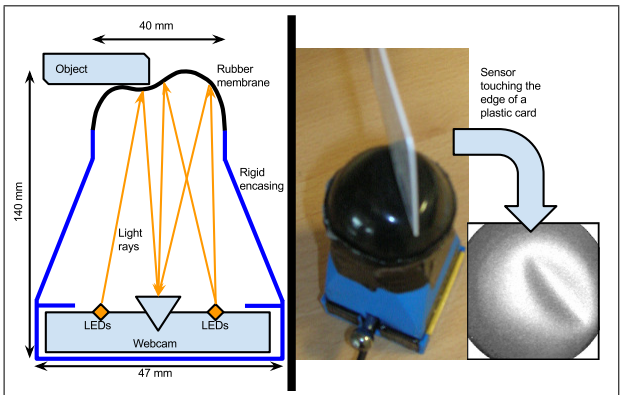


Fig. 2. The new tactile sensor design (left). The main body is 3D printed in ABS. The tip is a 1mm thick silicone rubber hemisphere. At the base (not visible) there is a USB eSecure web-cam (running at 640 by 480 pixels) with 8 LEDs illuminating the inside of the silicone hemisphere. As the tip makes contact with an object, it deforms resulting in a specific shading pattern (right).

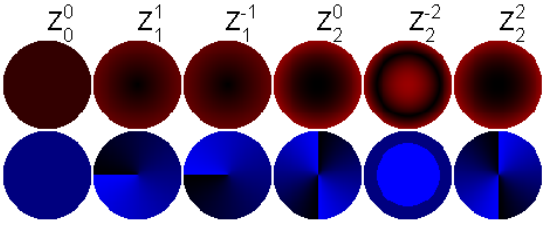


Fig. 3. The first Zernike polynomials evaluated on a unit disk. Here depicted as modulus (red) and phase (blue).

even and odd Zernike polynomials are defined as:

$$Z_n^m(\rho,\varphi) = R_n^m(\rho)\,\cos(m\varphi)$$

$$Z_n^{-m}(\rho,\varphi) = R_n^m(\rho)\,\sin(m\varphi),$$

Which can be indexed by:

$$Z_j = Z_{n(j)}^{m(j)}$$

Where $m(j), n(j)$ are Noll's indices of Zernike polynomials [24], and

$$R_n^m(\rho) = \sum_{k=0}^{(n-m)/2} \frac{(-1)^k\,(n-k)!}{k!\,\left(\frac{n+m}{2-k}\right)!\,\left(\frac{n-m}{2-k}\right)!}\,\rho^{n-2\,k}$$

Then, the d^{th} Zernike Moment of an image M is given by:

$$Zer_d(M) = \left| \sum_{i,j\in\{i^2+j^2\le n^2/2\}} M(i,j) Z'_d(i,j) \right|$$

Where,

$$Z'_d(i,j) := Z_j\left(\frac{\sqrt{(i^2+j^2)}}{\frac{\sqrt{2}}{2}n}, arctan\left(\frac{j-n/2}{i-n/2}\right)\right)$$

Once the Zernike moments are obtained from the entire training set, PCA is performed. The Zernike moments of images obtained during validation/testing are multiplied by the PCA dimensionality reduction matrix obtained during training. This process is hereafter referred to as "finding the Zernike-PCA moments". The number of components to be used is decided by inspecting the eigenvalues and retaining sufficiently many principal components so as to explain 95% of the variance in the training data.

IV. Object Learning and Recognition

The proposed model stores the Zernike-PCA moments of all tactile images and their corresponding object labels given during training. During testing, the Zernike-PCA moments of each new tactile image is compared against those stored values, and the likelihood of the new image, given each learnt object, is computed. This likelihood is defined as the normalized sum of n_C Normal probability density functions, where n_C is the number of training images used for object C. Each one of these is evaluated at the sensed image's Zernike-PCA value, centered at one of the training points, and with covariance given by the covariance matrix of all training points[2]. The process is depicted in Fig. 4.

Formally, let the training set be $X_C = \{X_{C,i}, i = 1, ..., n_C\}$, where $X_{C,i}$ is the Zernike-PCA moment vector corresponding to the i^{th} tactile image of object C, which was observed n_C times during training. Let W be the covariance matrix of X_C. Let $Y = \{Y_j, j = 1, ..., m\}$ be the sequence of Zernike-PCA moments (PCA reduction is performed using the dimensionality reduction matrix obtained from the training data), where Y_j represents the Zernike-PCA moments of the j^{th} tactile image of the object being sensed for recognition. Then the likelihood of Y_j for a given object class C is defined as:

$$P(Y_j|C) = \frac{1}{n_C}\sum_{i=1}^{n_C} \mathcal{N}(Y_i|X_{C,i}, W) \tag{1}$$

Where,

$$\mathcal{N}(Y_i|X_{C,i}, W) = \frac{e^{-\frac{1}{2}(Y_j - X_{C,i})^T W^{-1}(Y_j - X_{C,i})}}{\sqrt{\|W\|(2\pi)^d}}$$

Here, d is the dimensionality of the feature vector. Using this likelihood function a Naive Bayes classifier was implemented. This assumes that observed Zernike-PCA moments are statically independent. Note that PCA projection here helps to mitigate against correlations between features.

$$P(C|Y) = \alpha \prod_{j=1}^{m} P(Y_j|C)P(C)$$

Where α is just a normalizing constant, and $P(C)$ can be estimated from the number of times each object is observed during training, which in our case forms a uniform prior distribution. Therefore object recognition can be performed using maximum a posteriori:

$$C_{pred} = \underset{C}{\operatorname{argmax}}\, P(Y_j|C)$$

The computational complexity arises from Equation 1. Assuming there are n observations times during training, the complexity is $O(dn^2)$ during training and $O(d^2n)$ during testing.

V. Experiments and Results

Two experiments were performed to test the accuracy of the object recognition method outlined above: one to recognise objects seen before within a fixed collection, the other to test generalisation to unseen objects. Both experiments were carried out under the same setup.

A. Experimental setup

The system consisted of a 6 degrees of freedom (DOF) KUKA KR5-sixx-R650 robotic arm, a 6 DOF force-torque sensor mounted on its end effector, and the new tactile sensor mounted on the force-torque sensor (see Fig. 1). The force-torque sensor was used to detect touch events and to ensure the safety of the robot-object interaction.

[2]In practice, this is the diagonal matrix of variances, since X_C is the scores matrix resulting from PCA.

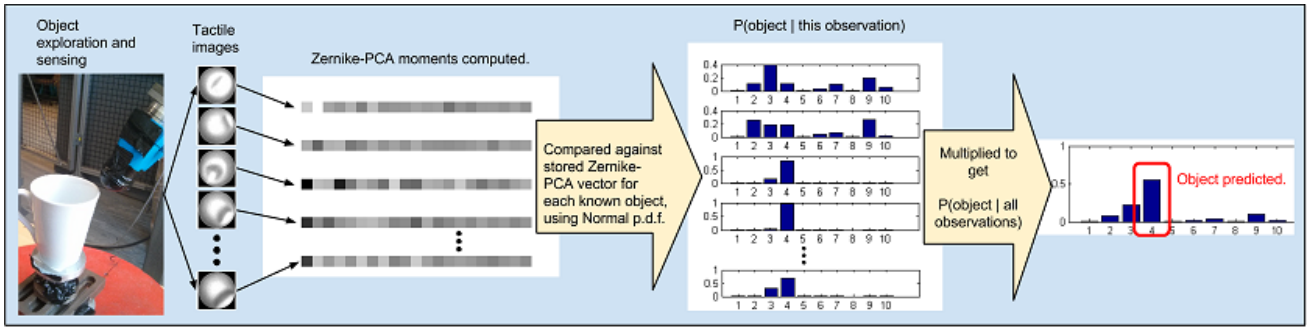


Fig. 4. The recognition process: from touch to object posterior probabilities.

The initial location of the object is assumed to be known, but its orientation is unknown. Limited unintentional pose alteration (less than 5% of object size) does occur during the experiments, as a consequence of contact. The aim is to have the robotic arm move the sensor to various point on the object surface and collect the tactile information autonomously. Each object was manually placed and secured in this location. The robotic arms is programmed to perform the following exploration procedure:

1) Define a "safety hemisphere" of radius 30cm about the assumed object centre. The hemisphere occupies the space above the object.
2) Generate a set of random points on that hemisphere.
3) Take the sensor to the next unvisited position in the list, facing inwards towards the centre point.
4) Move the sensor linearly inwards, until a normal force of 75 grams is detected.
5) Record the tactile image.
6) Retract the sensor linearly away from the object back to the imaginary sphere.
7) Back to step 3.

B. Object recognition

The objective of the first experiment was to automatically explore, learn and recognise objects from a set of 10 household objects (see Fig. 5): stapler, toothbrush, porridge pot, mug, shampoo bottle, box, pen, ball, textbook, water bottle (empty).

A total of 120 tactile images were collected for each object. These were split into 60 for training, 30 for validation and 30 for testing. A number of tests were attempted using the validation data set for testing. Initially, a Naive Bayes classifier using clustering was implemented, which resulted in approximately 70% accuracy after 30 touches, using k-means. Alternative clustering methods were tested, but did not improve performance. In particular Gaussian Mixture Models seemed suitable due to the natural representation of the likelihood function for observed data, but the parameter estimation led to an under-determined system for such a small data set. The final choice of inference system is non-parametric, and as such there is no need for a validation data set for parameter estimation. Of the 90 samples (training and testing) for each object, 100 different partitions (60 training images and 30 testing images) were made, the accuracy reported is the percentage of correct recognitions, averaged of these 100 iterations. Fig. 6 shows the confusion matrix after 5 and 15 touches.

After 15 touches the overall accuracy is 87% ; however, there is still a marked (approx. 19%) confusion between the toothbrush and the pen. These objects are very similar to touch in many of their local patches. This confusion represents 2.7% of the inaccurate predictions after 30 touches. There is high uncertainty about the stapler in the first 5 touches, perhaps reflecting the varied tactile features of its surface.

Fig. 7 shows the average accuracy for all objects, over 100 trials. As a comparison, best previous results (averaged over 7 trials) are shown [13]. The recognition accuracy follows a similar pattern in all methods, however our system gains a clear advantage from the start, and it stabilizes after about 25 touches.

C. Classifying unseen objects

In the second experiment, the potential for classification of previously unseen objects was preliminarily tested. The aim was to discern if the system had potential to classify objects that had not been used in training. Five previously

Fig. 5. The objects to be learnt and recognised.

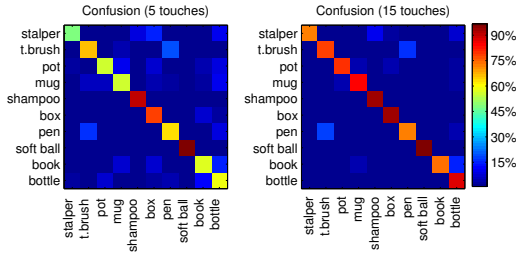


Fig. 6. Confusion matrices showing recognition errors for the 10 household objects, after 5 and 15 touches.

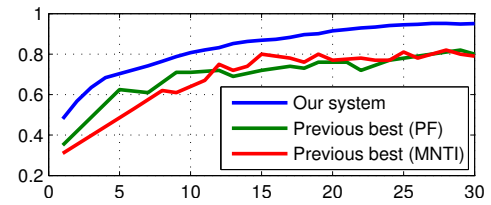


Fig. 7. Accuracy as a function of the number of touches, averaged over 100 trials. Comparison to previous work by [13], showing their best performing features: "Polar Fourier" (PF) and "Moment Normalized Translation Invariant" (MNTI).

untouched objects were sensed and attempted to be classified using the system outlined above. The objects used were: a plastic card, a different mug, a different pen, a smaller and harder ball, and another textbook (soft-back). This time the full data set for the 10 known objects was used for training, and 120 images of the unseen object were used in testing. Fig. 8 shows the posterior probabilities of each of the known 10 objects, assigned to each of the new objects, against the number of touches.

The plastic card is very different to any known objects and as such causes high confusion initially. The system finally settles for classifying it as a mug or a pot. The new pen is initially very confidently classified as a pen, but after 10 touches there is growing confusion with the pot model. This may be due to the rounded edge of the pot having a similar curvature to the pen. The other three objects are on average "correctly" classified. There is some confusion between the mug and the pot when classifying the new mug, which is understandable due to the similarity between the two known objects. These preliminary results show promise that the system may be generalisable to unseen objects, but are modest in scale and as such not conclusive: further research is required. It seems that objects very similar to the known ones (new book, new ball, new mug, new pen) are classified "correctly" very quickly, and as such the level of uncertainty at the beginning of the exploration could be used to inform a system that predicts new classes.

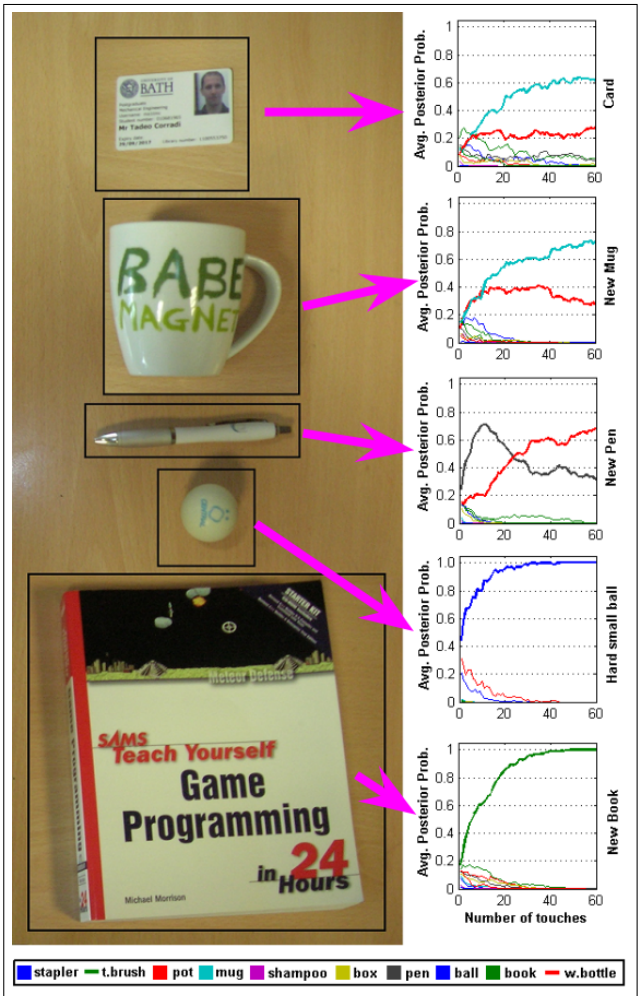


Fig. 8. Attempting to classify previously untouched objects. Posterior distribution over the known object classes, when testing is performed on five objects not sensed during training. Averaged over 100 trials.

D. Timings

All timings provided are for single-threaded, unoptimized, MATLAB code, running on a Core i7-4700MQ 2.4Ghz with 8Gb DDR3-1600 RAM. Zernike moment calculation took on an average of 3.7×10^{-3} s per tactile image. Feature dimensionality was always 21 or 22. For the first experiment (600 images in training), training took an average of 1.7×10^{-8} s, and testing 8.6×10^{-4} s per tactile image. For the second experiment (1200 images in training), training took an average of 1.7×10^{-8} s, and testing 1.2×10^{-3} s per tactile image. All these timings are substantially lower than the average time it takes the robotic arm to take a reading (approximately 30 seconds).

VI. Conclusion

A new inexpensive tactile sensor combined with an automated simple Bayesian object identity inference system were presented. They were shown to achieve accuracy in recognition outperforming state-of-the-art, for single contact, local appearance based tactile object recognition. The sensor

was made open source and can has a total material cost of approximately USD 30, substantially less than any other commercial or open source tactile sensor available, making it widely available to experts and hobbyists. A system was designed to autonomously collect tactile information from a range of household objects, using this new sensor, mounted on a robotic arm and aided by a force-torque sensor. These results are obtained using a very limited number of training, validation and testing images, about a third of previous similar work. In addition, preliminary results show potential for unseen object classification, yet more research is needed. Recognition is performed in real time.

Inference is performed using a Naive Bayes classifier. As such, there is a an assumption of independence between observed features. This assumption is potentially limiting and a more sophisticated probabilistic model may be needed as the number of classes grows larger.

At present, exploration takes approximately 30 seconds per reading, 30 minutes to learn an object's representation and 15 minutes to recognise it with 95% confidence. Whilst attempts were made to create a reactive system, robot control is relatively rigid. It would be interesting to explore ways of using machine learning to make the robot control more efficient and self-adapting. Future work will also include sensor fusion, attempting to harness the potential shown here to complement active vision systems.

ACKNOWLEDGMENT

We would like to thank the Bristol Robotics Laboratory for lending us the TacTip sensor.

REFERENCES

[1] T. Corradi, P. Hall, and P. Iravani, "Tactile Features: Recognising Touch Sensations with a Novel and Inexpensive Tactile Sensor," in *Advances in Autonomous Robotics Systems*, ser. Lecture Notes in Computer Science, M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish, Eds. Springer International Publishing, Jan. 2014, no. 8717, pp. 163–172.

[2] R. Dahiya, P. Mittendorfer, M. Valle, G. Cheng, and V. Lumelsky, "Directions Toward Effective Utilization of Tactile Skin: A Review," *IEEE Sensors Journal*, vol. 13, no. 11, pp. 4121–4138, 2013.

[3] N. Wettels, V. J. Santos, R. S. Johansson, and G. E. Loeb, "Biomimetic Tactile Sensor Array," *Advanced Robotics*, vol. 22, no. 8, pp. 829–849, 2008.

[4] K. Weiss and H. Worn, "The working principle of resistive tactile sensor cells," in *Mechatronics and Automation, 2005 IEEE International Conference*, vol. 1, 2005, pp. 471–476 Vol. 1.

[5] Pressure Profile Systems, "DigiTact II." [Online]. Available: http://www.pressureprofile.com/digitacts-sensors

[6] L. Jentoft, Y. Tenzer, D. Vogt, J. Liu, R. Wood, and R. Howe, "Flexible, stretchable tactile arrays from MEMS barometers," in *2013 16th International Conference on Advanced Robotics (ICAR)*, Nov. 2013, pp. 1–6.

[7] C. Chorley, C. Melhuish, T. Pipe, and J. Rossiter, "Development of a tactile sensor based on biologically inspired edge encoding," in *International Conference on Advanced Robotics, 2009. ICAR 2009*, 2009, pp. 1–6.

[8] S. Navarro, N. Gorges, H. Worn, J. Schill, T. Asfour, and R. Dillmann, "Haptic object recognition for multi-fingered robot hands," in *2012 IEEE Haptics Symposium (HAPTICS)*, 2012, pp. 497–502.

[9] H. Soh, Y. Su, and Y. Demiris, "Online spatio-temporal Gaussian process experts with application to tactile classification," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 4489–4496.

[10] M. Madry, L. Bo, D. Kragic, and D. Fox, "ST-HMP: Unsupervised spatio-temporal feature learning for tactile data," in *IEEE International Conference on Robotics and Automation (ICRA)(to appear)*, 2014.

[11] J. K. Kim, J. W. Wee, and C. H. Lee, "Sensor fusion system for improving the recognition of 3D object," in *2004 IEEE Conference on Cybernetics and Intelligent Systems*, vol. 2, 2004, pp. 1207–1212.

[12] N. Gorges, S. Navarro, D. Goğer, and H. Worn, "Haptic object recognition using passive joints and haptic key features," in *2010 IEEE International Conference on Robotics and Automation (ICRA)*, 2010, pp. 2349–2355.

[13] Z. Pezzementi, E. Plaku, C. Reyda, and G. Hager, "Tactile-Object Recognition From Appearance Information," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 473–487, 2011.

[14] A. Bierbaum, K. Welke, D. Burger, T. Asfour, and R. Dillmann, "Haptic exploration for 3D shape reconstruction using five-finger hands," in *2007 7th IEEE-RAS International Conference on Humanoid Robots*, 2007, pp. 616–621.

[15] N. Gorges, P. Fritz, and H. Wörn, "Haptic Object Exploration Using Attention Cubes," in *KI 2010: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, R. Dillmann, J. Beyerer, U. D. Hanebeck, and T. Schultz, Eds. Springer Berlin Heidelberg, Jan. 2010, no. 6359, pp. 349–357.

[16] N. Gorges, S. Navarro, and H. Worn, "Haptic object recognition using statistical point cloud features," in *2011 15th International Conference on Advanced Robotics (ICAR)*, 2011, pp. 15–20.

[17] A. Aggarwal, P. Kampmann, J. Lemburg, and F. Kirchner, "Haptic Object Recognition in Underwater and Deep-sea Environments," *Journal of Field Robotics*, Aug. 2014.

[18] M. Meier, M. Schopfer, R. Haschke, and H. Ritter, "A Probabilistic Approach to Tactile Shape Reconstruction," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 630–635, June 2011.

[19] M. Jin, H. Gu, S. Fan, Y. Zhang, and H. Liu, "Object shape recognition approach for sparse point clouds from tactile exploration," in *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec. 2013, pp. 558–562.

[20] R. A. Russell, "Object recognition by a 'smart' tactile sensor," in *Proceedings of the Australian Conference on Robotics and Automation*, 2000, pp. 93–8.

[21] A. Schneider, J. Sturm, C. Stachniss, M. Reisert, H. Burkhardt, and W. Burgard, "Object identification with tactile sensors using bag-of-features," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009*, 2009, pp. 243–248.

[22] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 489–497, May 1990.

[23] Q. Wu and P. Hall, "Prime Shapes in Natural Images." British Machine Vision Association, 2012, pp. 45.1–45.12.

[24] R. J. Noll, "Zernike polynomials and atmospheric turbulence," *Journal of the Optical Society of America*, vol. 66, no. 3, pp. 207–211, Mar. 1976.

Chapter 5

Visuo-tactile object recognition

5.1 Motivation: from tactile recognition to visuo-tactile recognition

Chapters 3 and 4 corroborated that the simple sensor is capable of both shape and object recognition, the next step was to test the hypothesis that a probabilistic visuo-tactile fusion model would provide higher accuracy than individual modalities, and to find out under which conditions this would be most marked. Recall that the motivation for a fusion system stems from the belief that, for humans, object representations are multi-modal [67, 68], with efforts to attempt to combine these modalities in robotics stemming from the 1980s [3] . In considering multi-modal robotic perception, the aims of this stage of the project were to attempt to answer the following questions:

1. Does the BathTip tactile sensor provide information that can complement visual input?

2. Under what circumstances is this most marked?

3. What sort of fusion method is most effective and efficient (in terms of recognition of objects) to achieve this multi-modal object representation?

5.2 Summary: visuo-tactile models compared

A similar set of 10 household objects as the one described in Chapter 4 was used. The tactile model and the tactile data collection procedure remained the same.

The choice of visual model was guided by the following considerations: probabilistic output, simplicity, quick to implement, limited power. Reviewing the options covered in Section 2.3.3, bag-of-features [28] was identified as a potential approach. More sophisticated approaches were available and relatively easily deployable (e.g. convolutional neural networks, [120]), but the model devised by Csurka et al. [28] was already achieving such high performance as to dominate over the tactile model in some contexts. If one modality were allowed to dominate overmuch, multi-modal fusion would not be justified or desirable. In fact, the visual effectiveness would occasionally be so high, it inspired the idea of artificially impairing vision.

Three multi-modal fusion systems were compared:

1. A baseline heuristic model based on an average between the probability posteriors predicted by the visual model and the tactile model.

2. A nearest-neighbour system that concatenates visual and tactile feature vectors (replicating the work of Yang et al. [127]).

3. A proposed system based on the product of the posterior distribution of the tactile and the visual models.

The first approach is based on the assumption that how much a modality is 'trusted' (the weight parameter) is linearly dependent on the number of training samples for such a modality. There are some complications with this assumption. Additional training for a specific class in vision does not necessarily result in better vision performance for all classes. The linear assumption is also problematic: accuracy and consistency of a classifier need not improve linearly with the number of training samples. Finally, it is difficult to quantify whether one tactile training sample should be given the same importance as one visual training sample. For these reasons, this approach should only be considered as a baseline heuristic for comparison to the other two.

In order to evaluate and compare these systems, and to begin to answer the aforementioned questions, several experiments were carried out involving object recognition. In particular, attention was paid to the matter of learning efficiency (how to maximise accuracy while minimising the number training samples), using a novel metric to assess it.

5.3 Results: when does multi-modal sensing matter?

In all cases, vision and touch combined improved accuracy over either modality alone. Of the three models compared, the proposed posterior product model produced the best results. The improvement was most marked when neither modality dominates. Learning efficiency (accuracy versus number of training samples) was not higher in general but did show improvements when vision was artificially impaired.

5.4 Paper: Object recognition combining vision and touch

The visual model, visuo-tactile fusion system, experiments and results, were published in the Journal for Robotics and Biomimetics [26]. The Statement of Authorship Form and the paper are found next.

This declaration concerns the article entitled:									
Object recognition combining vision and touch									
Publication status (tick one)									
draft manuscript		**Submitted**		**In review**		**Accepted**		**Published**	✓
Publication details (reference)	T. Corradi, P. Hall, and P. Iravani. "Object recognition combining vision and touch", Robotics and Biomimetics, 4 (2017), p. 2.								
Candidate's contribution to the paper (detailed, and also given as a percentage).	The candidate contributed to/ considerably contributed to/predominantly executed the… Formulation of ideas: 70%. Formulated ideas for fusion, data pipeline, experiment, hypothesis, comparisons. Supervisors suggested graphical model. Design of methodology: 100%. Design of methodology, including data collection and programming completely by me. Various metrics incuding learning rate also my own making. Experimental work: 100%. I collected all data, programmed all methods, including comparison methods, and code for evaluation. Presentation of data in journal format: 95%. I decided structure, wrote all drafts, prepared all figures, submitted, and responded to reviews. Supervisors provided feedback on drafts and helped with publication process.								
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.								
Signed	Tarlos					**Date**	03/04/2018		

Object Recognition Combining Vision and Touch

Tadeo Corradi, Peter Hall and Pejman Iravani

Abstract— **This paper explores ways of combining vision and touch for the purpose of object recognition. In particular it focuses on scenarios when there are few tactile training samples (as these are usually costly to obtain) and when vision is artificially impaired. Whilst machine vision is a widely studied field, and machine touch has received some attention recently, the fusion of both modalities remains a relatively unexplored area. It has been suggested that, in the human brain, there exist shared multi-sensorial representations of objects. This provides robustness when one or more senses are absent or unreliable. Modern robotics systems can benefit from multi-sensorial input, in particular in contexts where one or more of the sensors performs poorly. In this paper, a recently proposed tactile recognition model was extended by integrating a simple vision system in three different ways: vector concatenation (vision feature vector and tactile feature vector), object label posterior averaging and object label posterior product. A comparison is drawn in terms of overall accuracy of recognition and in terms of how quickly (number of training samples) learning occurs. The conclusions reached are: (i) the most accurate system is 'posterior product', (ii) multi-modal recognition has higher accuracy to either modality alone if all visual and tactile training data is pooled together and, (iii) in the case of visual impairment, multi-modal recognition "learns faster", i.e. requires fewer training samples to achieve the same accuracy as either other modality.**

I. Keywords

Object recognition, Sensor fusion, Tactile sensors, Robotic vision.

II. List of Abbreviations

- PCA: Principal Component Analysis
- SVM: Support Vector Machine

III. Introduction

It seems evident that the presence of multiple sensors, capable of capturing complementary information about the environment, is a desirable feature of modern robots [18], [11]. Indeed, there are indications that humans use similar mechanisms to process sensory information from vision and touch and that memories are multi-sensorial in nature [19], [38], [20]. In the field of Machine Vision, Object Recognition has been so well understood that, in some cases, artificial systems have surpassed human accuracy [13]. Machine touch has also received a great deal of attention recently. Whilst most commonly focused on texture recognition [9], [15], [21], [33], substantial efforts have been made to design object recognition systems using touch [27], [34], [26]. The question of how these modalities are to be used in conjunction remains, however, largely unanswered. Early attempts involved building geometric models of objects [3]. More recently, the field has received a lot more attention, consistently showing that sensor fusion outperforms either modality alone [18], [14], [12], [40]. Only [18], [40] specifically consider object recognition with a direct fusion of touch and vision, and this is done with grasping approaches. In this paper, a complete sensor fusion model is proposed for vision and touch, demonstrating its potential in object recognition with a small number of training samples. Unlike the aforementioned studies, which use grasping, a single-touch approach is used here, using a biologically inspired tactile 'finger' (see Fig. 1). In particular, for the cases where

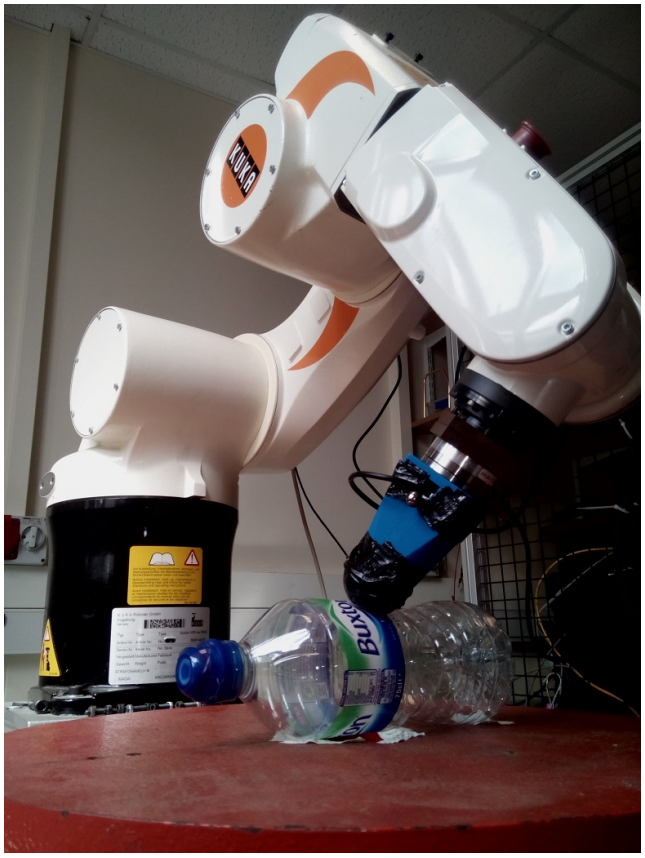


Fig. 1. Tactile data is collected autonomously by the tactile sensor developed in [7], mounted on a KUKA KR-650.

both modalities perform poorly independently (e.g. when vision is impaired), benefits are highlighted. It is also shown that, under certain conditions, the multi-modal systems are "faster learners" than vision and touch, i.e. they require fewer training samples to achieve comparable accuracy.

IV. Related Work

A. Tactile Object recognition

Kappassov et al. [16] distinguish between three types of tactile object recognition approaches: texture recognition, object identification (by which they mean using multiple tactile data types, such as temperature, pressure, etc. to identify objects based on their physical properties), and pattern recognition. This work falls within the last category. Most tactile recognition systems are based on recognition from grasping, i.e. using robotic hands or grippers equipped with multiple tactile sensors, where, often, the position of the fingers (proprioception) is also used as input. For example, using Self-Organising Maps and neural nets for household object recognition [27], using gaussian kernels to attain online learning of new objects [34], hierarchical feature learning (including temporal information) for object recognition [26], and multi-finger joint space sparse coding [22], all of which obtain near perfect accuracy. Recognition from grasping, however, requires the ability to grasp the object, whose identity is yet unknown, a non-trivial task. Alternatively, it is possible to recognise the object by means of individual contacts with a single tactile sensor. Some approaches involve volumetric reconstruction [10], [1] such as point-clouds or voxel space representation. Accuracy in these studies reaches 80% in some cases for 45 objects and only 10 touches, but 3D models of the objects are required in advance. Furthermore, there are technical challenges with scaling point-could and voxel representations. This paper focuses on this particular scope: single touch (non-grasping) object recognition. Schneider et al. [32] performed two-fingered grasps on a set of household objects, using a gripper equipped with tactile array sensors. From the resulting tactile images, a bag-of-tactile features approach was implemented to achieve over 84% accuracy in recognition. Their work uses information about the object relative position to the gripper. Pezzementi et al. [30] apply a predefined exploration routine with a single finger contact, to learn object models based on histograms of features (thus being the closest in data collection methodology to the work presented in this paper). Real object testing is limited to a set of 5 objects, achieving in excess of 90% accuracy for their best performing method. Recently, it was shown that single touch object recognition is possible even with a low resolution sensor [7]. Here,

that model is extended to account for visual information, comparing three different approaches to such multi-modal integration.

B. Visuo-tactile integration

Early attempts at integrating vision and touch were conducted by [3], using geometric models of objects and touch to complement unseen parts and again to estimate the parameters of a kinematic model for hand-object interactions [4]. Later, neural nets were used to fuse visual data and pressure data, showing that this sensor fusion was faster at learning and more accurate than either modality alone [18]. Recent work included fusion of RGB-D data and tactile data using an invariant extended Kalman filter to discover and refine 3D models of unseen objects [14]. It has been shown that fusion of vision and touch can be used to recognise the content of squeezed bottles [12], where the fusion of modalities outperforms either modality alone. Recently, Sun et al. [37] showed that sensing objects using vision and touch independently helps in identifications of suitable grasping plans. Visuo-tactile integration has also benefited the field of surface classification [36], where the variety of textures and patterns create difficulties for either modality alone. Most closely related to this paper are the works of Yang et al. [40] and of Liu et al. [23]. In [40], visuo-tactile integration shows great promise, demonstrating an improvement in accuracy using a simple weighted k-nearest-neighbour classifier to adjudicate a class label given vectors representing the tactile and visual input, obtaining a higher accuracy when both are combined rather than either used alone. [23] provides a visuo-tactile fusion model (using grasping) involving an innovative sparse coding algorithm for object instance recognition in a set of 18 objects, with similar results. This work is particularly impressive, as the sparse kernel encoding is robust to the inherently weak pairing between tactile and visual data. The

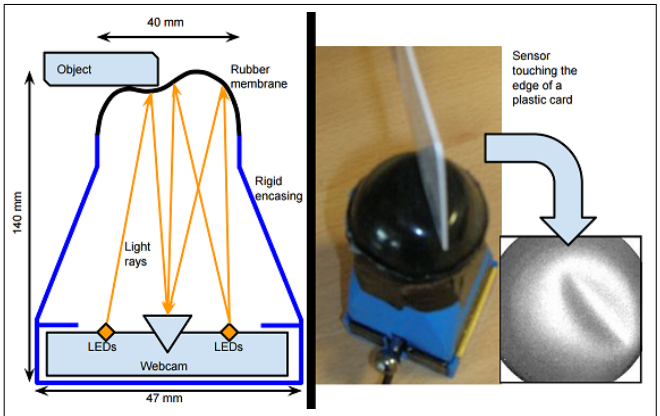


Fig. 2. The new tactile sensor design (left) first reported in [6]. The main body is 3D printed in ABS. The tip is a 1mm thick silicone rubber hemisphere. At the base (not visible) there is a USB eSecure webcam with 8 LEDs illuminating the inside of the silicone hemisphere. As the tip makes contact with an object, it deforms resulting in a specific shading pattern (right). Schematics and part details openly available at: https://github.com/Exhor/bathtip.

work presented in this paper contributes in four key aspects: a) Tactile data is collected with single touches (no grasping, no grippers) and the poses of the sensor and the object are ignored (no spatial information is used). b) Visual and tactile models developed are probabilistic, c) the main fusion model presented is both simple and grounded, and d) an analysis of arbitrarily impaired visual data is presented with a novel focus (learning efficiency).

V. Tactile and Visual models

A. Tactile model

The tactile sensor used here was first introduced in [6]. It comprises a camera inside a 3D-printed ABS enclosure, filming the shading pattern resulting from the deformation of an internally illuminated silicone rubber membrane, as it makes contact with an object (see Fig. 2). An extensive comparison of encodings and classifiers to best process the output of this sensor for shape and object recognition were covered in recent work [6], [7]. The algorithm devised in that work involves computing the Zernike moments [41] of a given normalised image (as read by the camera), and using PCA for dimensionality reduction. Zernike moments

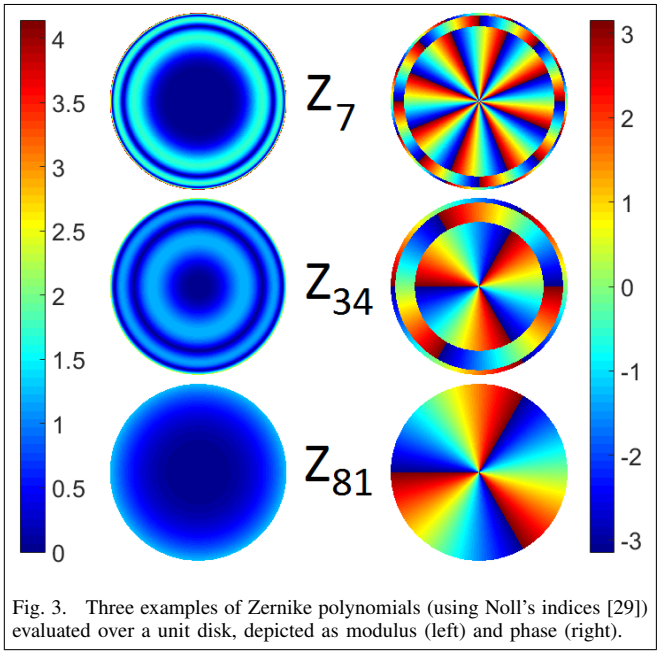


Fig. 3. Three examples of Zernike polynomials (using Noll's indices [29]) evaluated over a unit disk, depicted as modulus (left) and phase (right).

are obtained by computing the modulus of the inner product of Zernike polynomials (evaluated on a unit disk) with a given tactile image's intensity values (Fig. 3 shows a few sample Zernike polynomials). Using Zernike moments bears some immediate advantages: they provide a direct way of encoding images whose domain is the unit disk and they can provide rotational invariance [17] , which is ideal considering how the sensor works. Furthermore, they had already been used for basic visual shape recognition [39]. For more details, and comparisons to other encodings, see [7].

Each object is therefore represented by n vectors of size d, each containing the first d principal components of the Zernike-PCA descriptor of a tactile image captured during training. These n vectors are stored. A d-dimensional gaussian is centered at each one of these vectors, with covariance matrix obtained from the complete training dataset. The normalised sum of all these gaussians is the p.d.f. of the likelihood model, i.e. the model assigns a probability of observing a certain Zernike-PCA vector, for any given object: $P(tactile\_vector|object\_label)$.

Formally, let the training set of vectors be called $X_c = \{X_{c,i}, i = 1, ..., n\}$, were X_i is the Zernike-PCA moment vector the i^{th} tactile image of object c, which was observed n times during training.

Let W be the covariance matrix of X_c[1]. Let $t = \{t_j, j = 1, ..., m\}$ be the sequence of Zernike-PCA moments (where the PCA reduction is performed using the dimensionality reduction matrix obtained from the training data), where t_j represents the Zernike-PCA moments of the j^{th} tactile image of the object being sensed, and whose label is being predicted. Then, the likelihood of t_j for a given object label C is modelled as:

$$P(t_j|C) = \frac{1}{n_C} \sum_{i=1}^{n_C} \mathcal{N}(t_i|X_{C,i}, W)$$

Where,

$$\mathcal{N}(t_i|X_{C,i}, W) = \frac{e^{-\frac{1}{2}(t_j - X_{C,i})^T W^{-1} (t_j - X_{C,i})}}{\sqrt{\|W\|(2\pi)^d}}$$

Where d is the dimensionality of the feature vector. Assuming subsequent observations of the object are independent, and applying Bayes' Rule, the probability of each object label, C, given the set of observations t, is given by:

$$P(C|t) = \alpha \prod_{j=1}^{m} P(t_j|C)P(C) \quad (1)$$

Where α is a normalizing constant, and $P(C)$ can be estimated from the number of times each object is observed during training, which, in all cases covered here, forms a uniform prior distribution. Therefore, for touch-only recognition, object label inference is:

$$C_{touch} = arg\max_{C} P(C|t) \quad (2)$$

B. Visual model

The visual model is a simple bag-of-words model, using SURF [5] as features. K-means is used on the SURF de-

[1]In practice, this is very close to being the diagonal matrix of variances, since X_c is the scores matrix resulting from PCA.

scriptors of a pre-training dataset of unrelated images, for the purpose of dictionary creation. Each SURF feature descriptor of each object image is assigned a label (word), the closest k-means centre to it. Each image is thereafter represented by the histogram of these labels (words). During training, a one-vs-all r.b.f.-kernel support vector machine (SVM) is used on the normalised histograms corresponding to each object. During testing, a single visual image is used. The image's histogram is presented to all the SVMs, and a posterior distribution over object labels is computed using Platt scaling [31]. Specifically, let $s(v)$ be the score given by the SVM corresponding to label C to the visual histogram v of an object's image. Then the probability of label C is estimated as:

$$P(C|v) = \frac{1}{1 + exp(As(v) + B)} \tag{3}$$

Where A and B are two constants estimated by maximising the log likelihood of the training data (for details, see [31]). The predicted label for vision only is therefore:

$$C_{vision} = arg\max_{C} P(C|v) \tag{4}$$

VI. Visuo-tactile integration models

While attempting to integrate various modalities, recent work has focused in either deep learning and other neural approaches [35], [42], [28], probabilistic [24] or direct vector concatenation [40]. The first group has advantages in their ability to recognise relationship between input data at various levels of abstraction. However, they do require more data, which is a limitation in tactile robotics. In this paper, three approaches are compared, summarised in Fig. 4, and described below.

A. Posterior product

A straightforward approach to predicting an object label is to pick the label, C, that maximises the likelihood of observed data $P(v, t|C)$. Assuming conditional independence, $P(v, t|C) = P(v|C)P(t|C)$. Further assuming a uniform prior over class labels, applying Bayes' Rule and noting that $P(v)$ and $P(t)$ do not depend on C, means that maximising the product $P(v|C)P(t|C)$ over C is equivalent to maximising $P(C|v)P(C|t)$ over C.Therefore, the predicted label can be computed by:

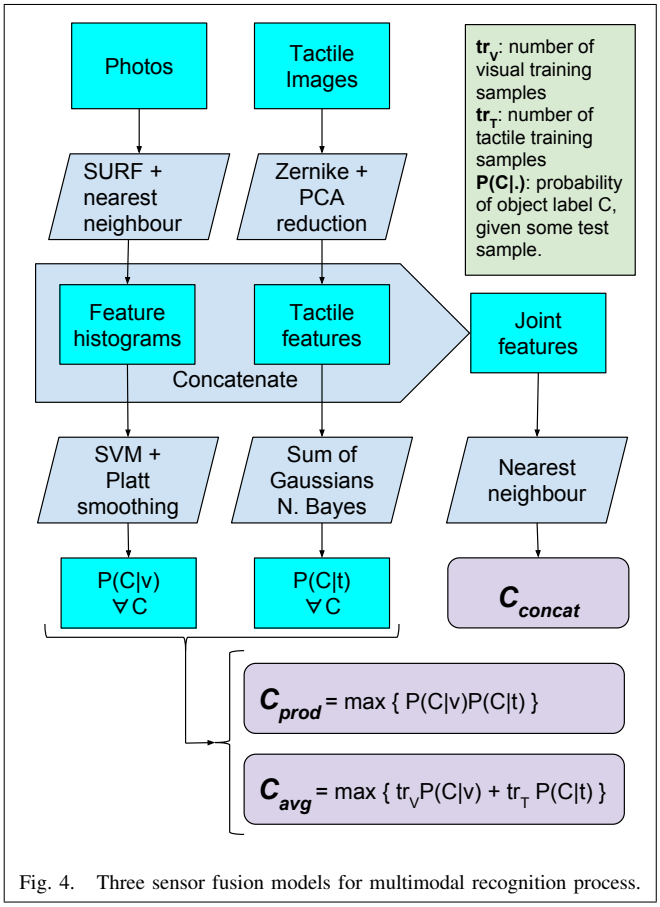


Fig. 4. Three sensor fusion models for multimodal recognition process.

$$C_{prod} = arg\max_{C}\{P(C|t)P(C|v)\} \tag{5}$$

Where $P(C|t)$ and $P(C|v)$ are the probabilities that the object being sensed has label C, given the tactile and the visual sensed data, respectively, as defined in equations (1) and (3). The assumption of independence in the above model is a simplification, since both vision and touch depend on the geometry of the object.

B. Vector concatenation

Similar to the work of [40], the second model presented directly concatenates the feature vectors for vision and touch and then label prediction is done by just finding the nearest neighbour in the training dataset. Nearest neighbour classification is known to be problematic in high-dimensional data [2], therefore, following the recommendations of [2], the $L_{0.1}$ distance metric is chosen. Thus, the label predicted is that for whom the distance to its closest training vectors is smallest. Let v_C is the nearest neighbour to a test image's histogram v of label C. Let $t_{C,1}, t_{C,2}, ..., t_{C,p}$ be the nearest tactile training vectors of label C to the testing vectors $t_1, t_2, ..., t_p$. Then, the predicted label for vector concatenation is:

$$C_{concat} = arg\min_{C} |v - v_C|_{L_{0.1}} + \frac{1}{p}\sum_{j=1}^{p} |t_j - t_{C,j}|_{L_{0.1}} \quad (6)$$

C. Weighted average of posteriors

A heuristic alternative is to consider the weighted average of posteriors, where the weight is set to the number of training samples for the modality. The rationale for such an approach is that the more experience (training samples) there is in a particular modality, the more it should influence a final decision. Thus, let tr_T and tr_V denote the number of training samples for a given simulation, then the predicted label for posterior average, C_{avg} given the input data, is given by:

$$C_{avg} = arg\max_{C}\{tr_T P(c|t) + tr_V P(c|v)\} \quad (7)$$

This approach would equate to vote counting, where both vision and touch cast votes for which class label should be chosen as most likely. The number of votes each casts being directly proportional to how many samples were used during their training.

Fig. 5. The 10 household objects used.

VII. Experiments and Results

Training was conducted on images of 10 objects (see Fig. 5) collected manually and tactile readings of the same objects, performed autonomously by a robot (illustrated in Fig. 1). The centre of the object was assumed to be known, then an angle of approach was chosen at random. The robot approached pointing the sensor inwards towards the assumed centre of the object, until there was a contact detected. A single image is retrieved from the sensor's camera and stored, before the arm retracts outwards and the process starts over (for more details, see [7]). The position and orientation of the sensor are not used, only the tactile images.

For some tests, vision was corrupted to produce "blotched" images to simulate visual impairment: images were covered by a small random number of randomly placed black circles occluding approximately 20% of the pixels. Images were resized to 300x300 pixels and set to gray-scale prior to processing. Some samples of unaltered and blotched images are depicted in Fig. 6.

Parameter estimation was performed on a validation subset of the data and the following optimal parameters were obtained:

Fig. 6. Sample of visual full images (top row), blotched images (bottom row). Blotches are in effect black, but are depicted orange for visibility.

- Number of principal components to retain in Zernike-PCA descriptors: 20
- Optimal feature descriptor from amongst SIFT [25], SURF, HOG [8]: SURF
- Size of the visual vocabulary for the SURF Bag-of-words model: 100

The remaining dataset was repeatedly split into training and testing subsets, each such split is referred to as a "simulation" (all data is from real robot experiments). The number of training samples varied in each simulation. During testing, visual posterior calculation is performed according to equation (3), with a single image. For tactile recognition, up to 30 tactile images were considered in sequence, to produce a tactile posterior calculation, as defined in equation (1). Notice that, at times, only a subset of the 30 tactile images was considered for testing. With these, C_{touch}, C_{vision}, C_{prod}, C_{concat} and C_{avg} were computed as defined in equations (2)-(7). Each simulation will produce one prediction per visual photo. Each photo will be randomly paired with up to 30 tactile images from the same object. Accuracy is defined as the mean average proportion of correct label predictions over all simulations. Let d be the number of simulations, assume each simulation has n_v testing photos, and let $y_{i,j}$ be the predicted label for an object whose true label is $x_{i,j}$, corresponding to the j^{th} photo of the i^{th} simulation, then the accuracy reported is

$$Accuracy = \frac{1}{d}\frac{1}{n_v}\sum_{i=1}^{d}\sum_{j=1}^{n_v}\mathbb{1}_{\{x_{i,j}\}}(y_{i,j}) \tag{8}$$

Where the label prediction $y_{i,j}$ is performed according to equations (2)-(7), and $\mathbb{1}$ is the indicator function.

Two experiments are reported. The first compared the accuracies of recognition of uni-modal and multi-modal approaches using all training data available. For the second experiment, the total number of training samples (visual plus tactile) is fixed a priori.

A. *Uni-modal and multi-modal recognition accuracy*

For the first experiment, 60 visual and 60 tactile training samples were used. Each simulation represents a different training/testing data split. A total of 700 simulations were run. As there are 10 objects, the baseline (random) recognition accuracy is 0.1.

During test time, for a given object, a single visual image was used for vision and a sequence of up to 15 tactile images corresponding to that object were used for touch. Fig. 7 shows mean accuracy as more and more tactile images were used at test time.

For the case of unaltered images (Fig 7, bottom), vision achieved 0.86 accuracy. For a single tactile image, touch only attained 0.43, whilst all multi-modal approaches provide an improvement over vision alone (albeit small). As more touches are used at test time, tactile accuracy obviously improves. As the gap in performance between the modalities narrowed, the relative improvement of multi-modal approaches seemed more marked.

For the case of blotched images (Fig 7, top), vision's accuracy is much lower at 0.5. When only one touch was allowed at test time, the tactile accuracy was still 0.43, and

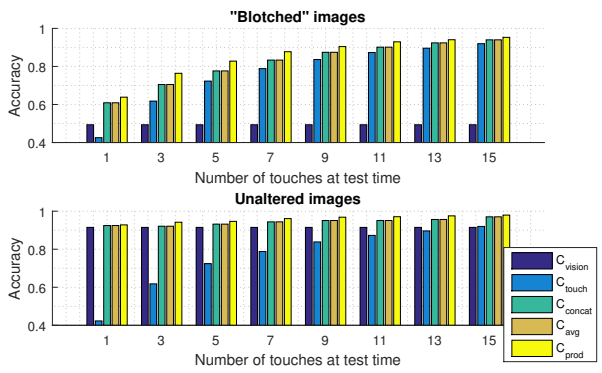


Fig. 7. Accuracy of recognition for 10 objects vs the number touches (tactile images) used at test time. Showing mean average over 700 simulations for each graph. Comparison of three approaches to multi-modal recognition.

the multi-modal approaches all showed a marked relative improvement. In this case, the accuracies of vision and touch started on a similar level, but touch evidently increased as more and more tactile images were used at test time. Even so, the multi-modal approaches showed an improvement over either modality in all cases.

In other words, the improvement in accuracy seemed smallest where the two modalities differed significantly in performance, and one dominated over the other. By contrast, when vision was impaired and few tactile images were allowed at test time, the improvement was most marked.

B. Learning efficiency: accuracy vs number of training samples

For the second experiment, the aim was to ascertain how efficient in terms of number of training samples the learning process was, with multi-modal representations, in comparison to each individual modality. The reasoning is that it may be considered "unfair" to compare a vision-only system which used 60 training samples against a visuo-tactile system that used 120 (60 visual and 60 touch). Instead, the total number of training samples was set to a fixed value and the accuracy for uni-modal and multi-modal were computed. For example, when the number of training samples was set to 40, tactile-only and visual-only recognition was performed using 40 training samples, but multi-modal recognition was performed using 20 visual and 20 tactile, or 35 visual and 5 tactile, or any other combination. This is different to all previous work encountered, where, when it comes to sensor fusion, all data from both modalities is typically used (such as in the first experiment).

At test time, a single image was used for vision, and a sequence of up to 30 tactile images for touch. Fig. 8 shows mean accuracy against total number of training samples. Following the findings in the first experiment, the reported number of tactile images used at test time was chosen so as to not allow either modality to dominate. That is, when "blotched" images were considered (top three graphs), only a few tactile images were needed for this purpose; but, in the case of full images (bottom three graphs), vision was stronger, so more tactile images were needed to achieve a similar degree of accuracy.

Consider the case of "unaltered" images, the lower part of Fig. 8. When 5 touches are allowed at test time (bottom left), vision is superior to touch. The accuracy of all multi-modal approaches fell short of vision's, namely it provides no improvement in this context. Even when 15 or 30 tactile images were used (bottom middle and bottom right), and there was no clear disparity in performance between vision and touch, the multi-modal approaches are not more "efficient" than one of the modalities alone, i.e. they require the same or more total training samples to achieve similar accuracy.

Now consider the case of using "blotched" images at test time (Fig. 8, top). When at least 40 training samples were used, the product of posteriors approach (C_{prod}) achieved higher accuracy than any other. As more touches were allowed at test time (top centre and right), the touch-only accuracy improved quickly, and the relative gain from multi-

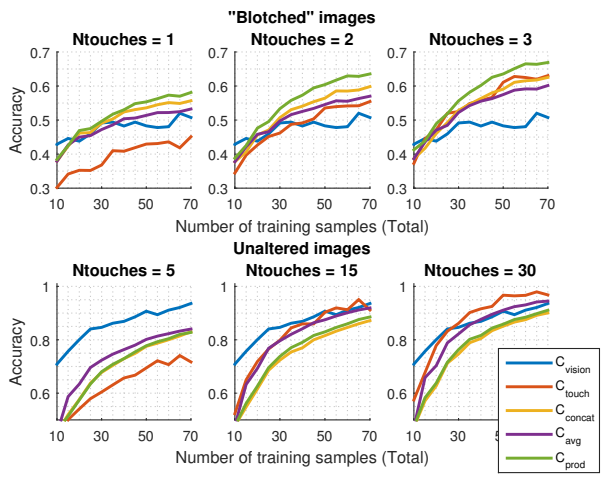


Fig. 8. Accuracy of recognition for 10 objects vs the number of training samples used. Showing mean average over 700 simulations for each graph. Comparison of the three approaches to multi-modal recognition. **"Ntouches"** stands for the number of tactile images used at test time.

modal approaches declined, to the point that only C_{prod} was visibly superior for the case of 3 touches at test time (top, right).

VIII. Conclusions and evaluation

A system was proposed for the purpose of visuo-tactile object recognition, by extending a recent tactile recognition model [7] and integrating it with a simple visual model. Three alternatives were considered for such integration, C_{concat}, C_{avg} and C_{prod}. Visuo-tactile approaches show considerable performance gains over either individual modality for the purpose of object recognition. In particular, the proposed method of posterior product outperforms both the weighted-average heuristic and the vector concatenation [40]. A novel comparison metric was proposed, fixing the total number of training samples a priori, so that, for example, a visuo-tactile approach using 30 visual and 30 touch training samples is compared to visual-only or tactile-only systems using 60 training samples. Under this new metric, the superiority of multimodal approaches (and of poterior-product in particular) was only found where vision was impaired artificially. It must be borne in mind that vision presents a remarkably high accuracy from very few training samples for unaltered images. Therefore, it is inherently more challenging to obtain improvements. This highlights a limitation of this metric, for there may be a fairer comparison. Even under such consideration, for "blotched" images, higher accuracy was obtained with N visual plus N tactile training samples, than 2N visual and than 2N tactile, for all models and values of $N > 20$. The artificially introduced visual impairment had the effect of overall lowering the accuracy of vision, and, where this was combined with lower accuracy from touch, the greatest improvement was obtained by the multi-modal approaches, in particular, by the product of posteriors, C_{prod}. Further work will explore the potential of these models for object class recognition and fine-grained recognition, using multiple instances of each class and thus the extension to a larger dataset.

IX. Declarations

A. Funding

The work reported here was funded by the Engineering and Physical Sciences Research Council (EPSRC), via a PhD scholarship awarded to T. Corradi. EPSRC approved the initial PhD proposal but had no direct input on either data collection, design or execution of the experiments reported, nor on the writing of the manuscript.

B. Competing interests

The authors declare that they have no competing interests.

C. Authors' contributions

PH proposed the zernike moment idea and the probabilistic approach. PI contributed in the design of the sensor, experiment design and evaluation. TC proposed the idea, carried out literature review, collected data, executed the experiments and analysis, and drafted the manuscript. All

authors reviewed the manuscript at various stages, including approval of its final version.

REFERENCES

[1] Achint Aggarwal, Peter Kampmann, Johannes Lemburg, and Frank Kirchner. Haptic object recognition in underwater and deep-sea environments. *Journal of Field Robotics*, 32(1):167–185, aug 2015.

[2] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. *Database Theory ICDT 2001*, pages 420–434, 2001.

[3] Peter K. Allen. Integrating vision and touch for object recognition tasks. *International Journal of Robotics Research*, pages 7:15–33, 1988.

[4] Peter K. Allen, Andrew T. Miller, Paul Y. Oh, and Brian S. Leibowitz. Integration of vision, force and tactile sensing for grasping. *International Journal of Intelligent Machines*, 4(October 2015):129–149, 1999.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3951 LNCS, pages 404–417. 2006.

[6] Tadeo Corradi, Peter Hall, and Pejman Iravani. Tactile features: Recognising touch sensations with a novel and inexpensive tactile sensor. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8717 LNAI, pages 163–172. Springer Verlag, 2014.

[7] Tadeo Corradi, Peter Hall, and Pejman Iravani. Bayesian tactile object recognition: Learning and recognising objects using a new inexpensive tactile sensor. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, volume 2015-June, pages 3909–3914. Institute of Electrical and Electronics Engineers Inc., 2015.

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume I, pages 886–893, 2005.

[9] Sergio Decherchi, Paolo Gastaldo, Ravinder S. Dahiya, Maurizio Valle, and Rodolfo Zunino. Tactile-Data Classification of Contact Materials Using Computational Intelligence. *IEEE Transactions on Robotics*, 27(3):635–639, jun 2011.

[10] Nicolas Gorges, Peter Fritz, and Heinz Woern. Haptic Object Exploration Using Attention Cubes. In Rüdiger Dillmann, Jürgen Beyerer, Uwe D. Hanebeck, and Tanja Schultz, editors, *Ki 2010: Advances in Artificial Intelligence*, volume 6359 of *Lecture Notes in Computer Science*, pages 349–357. Springer Berlin Heidelberg, jan 2010.

[11] Nicolas Gorges, Stefan Escaida Navarro, Dirk Göger, and Heinz Wörn. Haptic object recognition using passive joints and haptic key features. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 2349–2355, 2010.

[12] Puren Guler, Yasemin Bekiroglu, Xavi Gratal, Karl Pauwels, and Danica Kragic. What's in the container? Classifying object contents from vision and touch. In *IEEE International Conference on Intelligent Robots and Systems*, pages 3961–3968, sep 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034. IEEE, dec 2015.

[14] J. Ilonen, J. Bohg, and V. Kyrki. Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing. *The International Journal of Robotics Research*, 33(2):321–341, oct 2014.

[15] Nawid Jamali and Claude Sammut. Majority voting: Material classification by tactile sensing using surface texture. *IEEE Transactions on Robotics*, 27(3):508–521, 2011.

[16] Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands Review. *Robotics and Autonomous Systems*, 74(PA):195–220, dec 2015.

[17] Alireza Khotanzad and Yaw Hua Hong. Invariant Image Recognition by Zernike Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990.

[18] Ji Kyoung Kim, Jae Woo Wee, and Chong Ho Lee. Sensor fusion system for improving the recognition of 3D object. In *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, volume 2, pages 1207–1212, 2004.

[19] Simon Lacey, Christine Campbell, and K. Sathian. Vision and touch: Multiple or multisensory representations of objects? *Perception*, 36(10):1513–1521, 2007.

[20] Simon Lacey and K. Sathian. Visuo-haptic multisensory object recognition, categorization, and representation. *Frontiers in Psychology*, 5(JUL):730, 2014.

[21] Hongbin Liu, Xiaojing Song, Joao Bimbo, Lakmal Seneviratne, and Kaspar Althoefer. Surface material recognition through haptic exploration using an intelligent contact sensing finger. In *IEEE International Conference on Intelligent Robots and Systems*, pages 52–57, 2012.

[22] Huaping Liu, Di Guo, and Fuchun Sun. Object Recognition Using Tactile Measurements: Kernel Sparse Coding Methods. *IEEE Transactions on Instrumentation and Measurement*, 65(3):656–665, mar 2016.

[23] Huaping Liu, Yuanlong Yu, Fuchun Sun, and Jason Gu. Visual Tactile Fusion for Object Recognition. *IEEE Transactions on Automation Science and Engineering*, pages 1–13, 2016.

[24] Ming Liu, Lujia Wang, and Roland Siegwart. DP-Fusion: A generic

framework for online multi sensor recognition. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 7–12, 2012.

[25] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[26] Marianna Madry, Liefeng Bo, Danica Kragic, and Dieter Fox. ST-HMP: Unsupervised Spatio-Temporal feature learning for tactile data. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 2262–2269, 2014.

[27] Stefan Escaida Navarro, Nicolas Gorges, Heinz Wörn, Julian Schill, Tamim Asfour, and Rüdiger Dillmann. Haptic object recognition for multi-fingered robot hands. In *Haptics Symposium 2012, HAPTICS 2012 - Proceedings*, pages 497–502, 2012.

[28] Kuniaki Noda, Hiroaki Arie, Yuki Suga, and Tetsuya Ogata. Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems*, 62(6):721–736, nov 2014.

[29] Robert J. Noll. Zernike polynomials and atmospheric turbulence. *Journal of the Optical Society of America*, 66(3):207, mar 1976.

[30] Zachary Pezzementi, Erion Plaku, Caitlin Reyda, and Gregory D. Hager. Tactile-object recognition from appearance information. *IEEE Transactions on Robotics*, 27(3):473–487, 2011.

[31] J Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[32] A Schneider, J Sturm, C Stachniss, M Reisert, H Burkhardt, and W Burgard. Object identification with tactile sensors using bag-of-features. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on DOI - 10.1109/IROS.2009.5354648*, pages 243–248, 2009.

[33] Jivko Sinapov, Vladimir Sukhoy, Ritika Sahai, and Alexander Stoytchev. Vibrotactile recognition and categorization of surfaces by a humanoid robot. *IEEE Transactions on Robotics*, 27(3):488–497, 2011.

[34] Harold Soh, Yanyu Su, and Yiannis Demiris. Online spatio-temporal Gaussian process experts with application to tactile classification. In *IEEE International Conference on Intelligent Robots and Systems*, pages 4489–4496, 2012.

[35] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2222–2230. Curran Associates, Inc., 2012.

[36] Matti Strese, Clemens Schuwerk, Albert Iepure, and Eckehard Steinbach. Multimodal Feature-based Surface Material Classification. *IEEE Transactions on Haptics*, pages 1–1, 2016.

[37] Fuchun Sun, Chunfang Liu, Wenbing Huang, and Jianwei Zhang. Object Classification and Grasp Planning Using Visual and Tactile Sensing. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 1–11, 2016.

[38] Nivaldo Vasconcelos, Janaina Pantoja, Hindiael Belchior, Fábio Viegas Caixeta, Jean Faber, Marco Aurelio M Freire, Vinícius Rosa Cota, Edson Anibal de Macedo, Diego a Laplagne, Herman Martins Gomes, and Sidarta Ribeiro. Cross-modal responses in the primary visual cortex encode complex objects and correlate with tactile discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, 108(37):15408–15413, sep 2011.

[39] Qi Wu and Peter Hall. Prime Shapes in Natural Images. In *Procedings of the British Machine Vision Conference 2012*, pages 45.1–45.12. British Machine Vision Association, 2012.

[40] Jingwei Yang, Huaping Liu, Fuchun Sun, and Meng Gao. Object recognition using tactile and image information. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1746–1751. IEEE, dec 2015.

[41] F. Zernike. Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. *Physica*, 1(7-12):689–704, may 1934.

[42] Wenping Zhang and Hong Zhang. Online kernel-based multimodal similarity learning with application to image retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9227(iii):221–232, 2015.

Chapter 6

Bayesian Visuo-tactile object classification and instance recognition

6.1 Motivation: verifying scalability and attempting classification

Chapter 5 validated the viability of the fusion model for object recognition in a small data set. Questions remained about its scalability, i.e. whether it would continue to accurately recognise objects if the data set was enlarged. Furthermore, the conclusions of the work presented in Chapter 4 included the potential for tactile classification, that is, the ability to predict the known class (mug, bowl, bottle, etc.) of a new object, not present during training.

6.2 Summary: larger data set and object classification

With the aforementioned considerations in mind, the next stage of the project aimed to:

- Collect and make available the largest visuo-tactile household object database to date, including 60 objects, 6 of each class: shoe, can, box, bottle-empty, bottle-full, bowl, ball, mug, stapler, soft-toy.

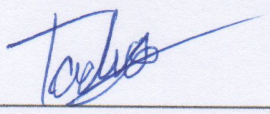
- Test the fusion model's scalability, i.e. its ability to recognise any of the 60 objects. This is rendered even more challenging by the close similarity between objects of the same class.
- Attempt tactile and visuo-tactile object classification. This has never been achieved to date.
- Explore potential real-life applications of the technology by means of a preliminary experiment: classification of 10 objects, submerged in murky water.

6.3 Results: object classification using touch and vision

For the first time in reported literature, tactile object classification was achieved, obtaining an accuracy of between 0.3 and 0.65, depending on the number of tactile images considered at test time. In all cases, accuracy was higher for the sensor fusion model, from 0.65 to 0.82 depending on the number of tactile images and whether or not the images were blotched. Once again, the largest improvements were seen were neither modality dominated (where their independent accuracies were close).

6.4 Paper: Bayesian object classification and instance recognition combining vision and touch

The database description, experiment details and results are currently in draft format and will soon be submitted for publication. The Statement of Authorship Form and the manuscript can be found next.

This declaration concerns the article entitled:												
Bayesian Visuo-tactile object classification and instance recognition.												
Publication status (tick one)												
draft manuscript	✓	**Submitted**		**In review**		**Accepted**		**Published**				
Publication details (reference)	N/A											
Candidate's contribution to the paper (detailed, and also given as a percentage).	The candidate contributed to/ considerably contributed to/predominantly executed the… Formulation of ideas: 85%. I proposed the idea of classification on a larger dataset, and proposed the experiment and analysis. Supervisors proposed idea of underwater test. Design of methodology: 85%. I proposed the machine learning pipeline, the experiment setup (dry experiment), and evaluation. Supervisors proposed details on underwater experiment using underwater cameras. Experimental work: 100%. I prepared all equipment, collected all data, programmed all code, including graphing and evaluation. Presentation of data in journal format: 95%. I structured the paper, wrote all drafts, produced all figures and other content. Supervisors gave feedback.											
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.											
Signed							**Date**	03/02/2018				

Bayesian object classification and instance recognition combining vision and touch

Tadeo Corradi, Peter Hall, Pejman Iravani

Abstract

The first example of tactile and visuo-tactile object class recognition is presented. The largest visuo-tactile household object database to date is made available, comprising 60 objects (10 classes, 6 instances). A Bayesian sensor fusion system involving vision and touch is deployed in both Object Class Recognition and Object Instance Recognition. Furthermore its potential is exemplified with an underwater object classification experiment. Recognising objects and object classes using multiple senses brings a number of benefits, including robustness to sensor failure, adverse conditions, ability to capture a wider range of object properties (e.g. stiffness, roughness). The model is validated by performing object instance recognition (identifying individual objects, e.g. "mug-04" in a database where other similar objects are present, such as mug-01, mug-02...) and object class recognition (correctly predicting an unseen object's class, e.g. "mug"). The results support sensor fusion as consistently more accurate in both problems (especially when vision is impaired). The database is made available so this baseline result can be improved.

1 Introduction

Combining multiple sensors, and thus perceiving a wider range of features, provides advantages for robotics systems [62, 45]. Humans are believed to use a multi-sensorial representation of objects for the purposes of object recognition [67, 119, 68]. While machine vision has been subject to substantial research (to the point that its accuracy is comparable or exceeding that of humans [50]), machine touch is less well understood. In part this is due to the lack of a standarised approach. Vision is largely standarised both in the sensors used and the format of the input, there is no such consensus in tactile robotics [29]. Tactile sensing

research focused initially mostly on texture classification [30, 55, 74, 107], and later on object recognition [85, 109, 81]. Combining vision and touch is still an open problem. Recent efforts show that the idea of multi-modal representations has merits in increasing recognition accuracy, with respect to either modality alone [62, 54, 49, 127], yet only [62, 127] consider the context of vision and touch. Recent work also shows that combining vision and touch has the potential to increase accuracy in recognition significantly if both modalities have low independent performance [26].

In this paper, visuo-tactile classification and instance recognition for a large object database are demonstrated. The first is defined as the ability to correctly predict the class of unseen objects (the object itself was not present during training, but other objects of its class were). The second is defined as the ability to correctly recognise a particular known object, where multiple similar objects exist in the database (the other objects of the same class).

The first problem has never been tackled using touch only nor using a fusion of vision and touch. The potential of the system for practical contexts is demonstrated with a further experiment involving classification of unseen objects submerged in murky water.

2 Related Work

In this work, two related problems are tackled: *instance recognition* (recognising an object which was sensed during training) and *object classification* (recognising the class of an object, where the object itself was not sensed during training, but other objects of the same class were).

2.1 Tactile Object Instance Recognition

The problem of Tactile Object Recognition is often tackled by means of grasping robotic hands or grippers, equipped with multiple tactile sensors of various types and configurations. Such configurations being the advantage of using (either explicitly or implicitly) information about the location of the sensors grasping an object (proprioception). For example, Self-Organising Maps and neural nets have been used for household object recognition [85] concatenating proprioception and tactile features. Gaussian Kernels have been designed to model the dynamic tactile sensations as perceived by a closing anthropomorphic hand, achieving on-line

learning of new objects [109], and recently [108] a variation thereof which is able to distinguish between full, half-full, and empty bottles. More recently, Hierarchical Feature Learning (including temporal information) has been used to learn tactile features in an unsupervised manner, again for the purpose of object recognition [81], obtaining near perfect accuracy. Simple features using pressure sensors only, when combined with proprioception, can achieve near perfect accuracy amongst 11 household objects some of which are very similar [110].

Recognition from grasping, however, requires the choice and configuration of a robotic hand, and the ability to grasp the -as of yet unknown- object, which is sometimes a complex problem. Instead, it is possible to perform object recognition using individual contacts with a single tactile sensor. This has been done with approaches that involve the 3D reconstruction of objects [44, 1, 121], using point-clouds or voxel space. These bring other complications, such as difficulty with scaling to large databases, and the computational complexity of volumetric registration/matching. Recently, a mixed approach was proposed which combines point clouds with feature-based recognition, achieving excellent results for tactile object recognition [79]. An alternative to volumetric approaches are bag-of-features methods, i.e. those which discard the geometric information (the location of the sensor during contact), and merely consider the tactile features extracted. One example is the work of Pezzementi et al. [91], which uses simulations to compare various methods of feature extraction, obtaining close-to-perfect recognition accuracy in a small set of objects. Drimus et al. [31] use tactile images' pixel intensity mean and standard deviation as features in a time series (dynamic touch), compared using dynamic time warping, over 10 objects, achieving in excess of 90% accuracy with a single sensor. Recently, it has been shown that bag-of-feature approaches are capable of tactile-only object recognition [25].

2.2 Visuo-tactile Instance Recognition

In the 1980s, pioneering the field of visuo-tactile integration, Allen [2] used geometric models of objects and used touch to fill in the invisible parts of objects. Later, his work was extended to estimate the parameters of a kinematic model for hand-object interactions [4], again combining vision and touch. In the 2000s, artificial neural nets were designed to combine visual input with pressure (one-dimensional tactile) input, displaying a faster learning cycle for the sensor-fusion model when compared to either modality alone [62]. More recently, Ilonen et al.

[54] have shown that Invariant Extended Kalman Filters can be used to fuse vision and touch to incrementally refine the 3D model of an unknown object. This same ideas has been realised by Bjorkman et al. [12] using Gaussian Processes over Zernike and curvature features. Guler et al. [49] conducted an experiment where the combination of vision and touch outperforms the independent modalities when recognising the contents of bottles by squeezing them. Combining vision and touch can also be of use for the purpose of planning grasps of unknown objects by means of classifying them into broad categories [113]. Yang et al. [127] combine vision and touch by means of concatenating the feature vectors extracted with each modality and use a nearest neighbour classifier with a weighted distance metric. In their work there are 18 objects, each represented by 10 photos and 10 grasp-touch sequences, using tactile sensors placed on the fingers of a 3-finger hand). Many of the objects can be considered very similar to one another (different sized cylinders, different coloured bottles). Recent work [26] shows the advantages of sensor-fusion for Object Instance Recognition with a database of 10 distinct objects, especially when both sensors were performing poorly independently. Recently, a visuo-tactile fusion model (using grasping) involving an innovative sparse coding algorithm for object instance recognition has been shown to achieve high accuracy in a set of 18 objects split over 5 classes, where most of the confusion in recognition arose within-class [73].

2.3 Visuo-tactile Object Class Recognition

Interest is growing in exploring multi-sensory object representations [61] and soon we may see the first large-scale visuo-tactile database [14]. A tactile-only attempt at object classification is reported by Gu et al. [48], who focus on shape recognition to distinguish between cuboids, cones, spheres, prisms and cylinders. This is a form of class recognition based on geometry. They capture point-clouds of 30 objects (6 for each shape) and use k-means clustering and random forests for classification, obtaining an accuracy of 87%. Similarly, in [112], context specific feature extraction and fusion to recognise materials is presented achieving 74% when combining all modalities, higher than any subset of them.

A closely related problem is that of binary adjective prediction [22] (e.g. determining if an object is smooth, coarse, soft...). For example, Gao et al. [41] use deep learning (one net for vision and one for touch, plus a fusion layer) to obtain state-of-the-art results over 24 adjectives in the PHAC-2 database [22]. Adjective

prediction can be interpreted as a form of non-exclusive binary classification.

The closest work to this paper is that of Sanchez-Fibla et al. [102], which combines visual and tactile information for curvature prediction; their work hints at its potential for object categorization, yet does not explore this fully.

This paper presents the first example of tactile object class recognition and visuo-tactile class recognition, as well as making public the largest visuo-tactile household object database to date (10 classes, 6 instances per class, totaling 60 objects). The Bayesian sensor-fusion model is further validated by performing object instance recognition within the database.

Furthermore, a proof-of-concept for class recognition of objects in a realistic context is presented: classification of objects submerged under murky water. For each of these scenarios, a comparison is drawn between the multi-modal system and each modality alone, highlighting accuracy gains.

3 Tactile and Visual models

This section describes how tactile and visual data are processed and stored during training and how they are combined to obtain probabilities for each known class of objects during testing. The recognition pipeline for tactile and visual input are showing in Fig. 6-1 and Fig. 6-2, respectively. Where instance recognition is being attempted, an object "label" refers to its instance (e.g. mug-4), while if class recognition is being attempted, the object label refers to its class (e.g. "mug").

3.1 Tactile model

The tactile object model used here was first introduced in [24]. It involves capturing tactile images using an optics-based tactile sensor [24], which takes photos of a deformable rubber membrane as it makes contact with an object (See Fig. 6-3).

Such images are then reduced in dimensionality using Zernike moments [132] and Principal Component Analysis, resulting in a vector of size 20. So, if during training, 54 tactile images are used to learn the representation of an object, the object model is a matrix of size 54 by 20.

Formally, for each object of label, c, let the training set of vectors be $X_c = \{X_{c,1}, X_{c,2}, ...X_{c,n_c}\}$, were $X_{c,i}$ is the Zernike-PCA moment vector the i^{th} tactile

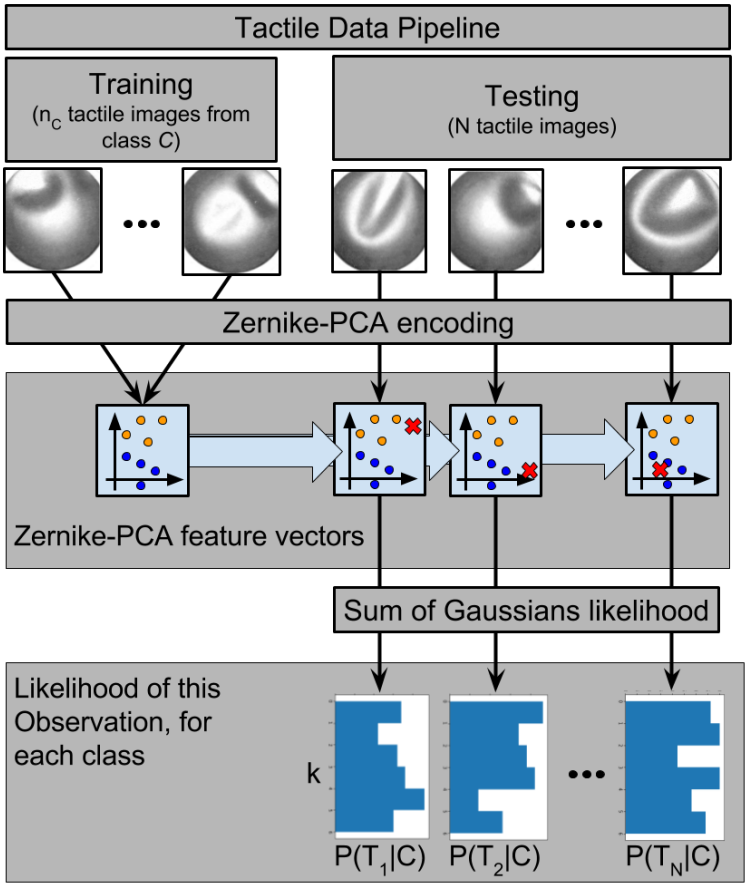


Figure 6-1: *Tactile data likelihood model. First proposed by [25].*

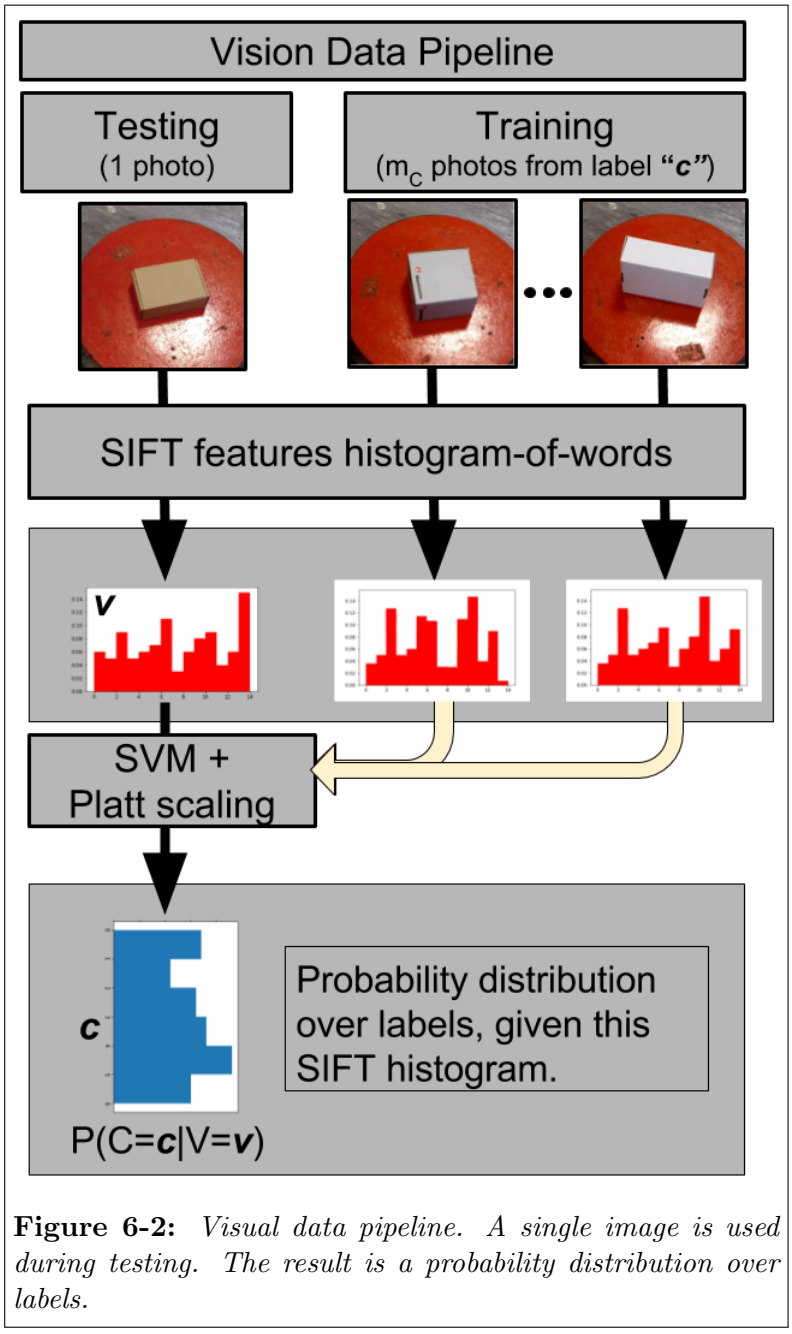


Figure 6-2: *Visual data pipeline. A single image is used during testing. The result is a probability distribution over labels.*

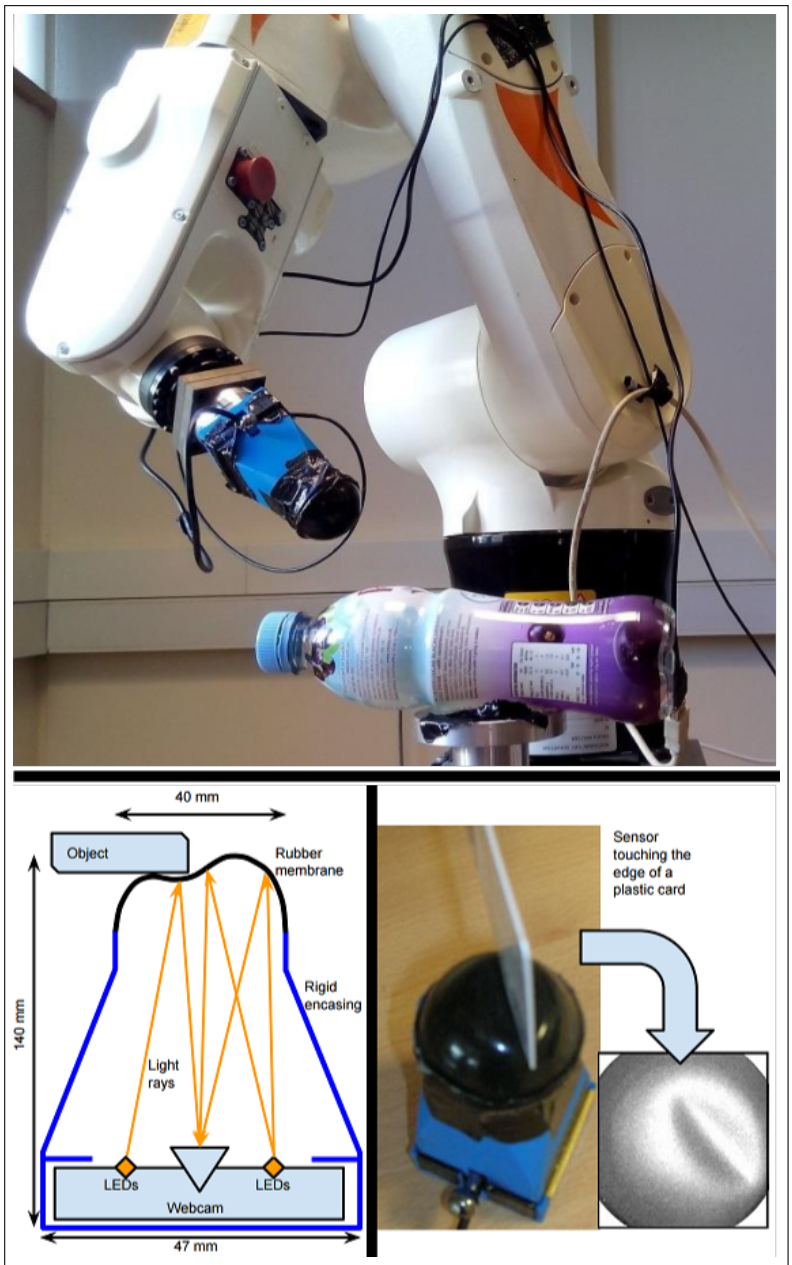


Figure 6-3: *The tactile sensor, mounted on a kuka robotic arm, automatically records tactile sensations at random contact points in each of the 60 objects. The tactile sensor used (bottom, left), first reported in [24]. The main body is 3D printed in ABS. The tip is a 1mm thick silicone rubber hemisphere. At the base (not visible) there is a USB web-cam with 8 LEDs illuminating the inside of the silicone hemisphere. As the tip makes contact with an object, it deforms resulting in a specific shading pattern (bottom, right). Schematics and part details are available at: https://github.com/Exhor/bathtip.*

image, which was observed n_c times during training. Let W be the covariance matrix of X_c.

During testing, Let $t = \{t_1, t_2, ..., t_N\}$ be the sequence of Zernike-PCA moments of the N tactile images of the object being sensed (PCA reduction is performed using the dimensionality reduction matrix obtained from the training data), and whose label (instance or class) is being predicted. The marginal likelihood of the observed tactile vector, t_k, given the object label, c, is modelled by:

$$P(T = t_k|C = c) = \frac{1}{n_C}\sum_{i=1}^{n_C}\mathcal{N}(t_k|X_{c,i}, W)$$

That is, a sum of Gaussian densities centered at the training vectors ($X_{c,i}$), with covariance determined by the covariance of the training vectors, evaluated at the testing vector, t_k,

$$\mathcal{N}(t_k|X_{c,i}, W) := \frac{exp(-\frac{1}{2}(t_k - X_{c,i})^T W^{-1}(t_k - X_{c,i}))}{\sqrt{\|W\|(2\pi)^d}}$$

Here, d is the dimensionality of the feature vector ($d = 20$). Assuming subsequent observations of the object are independent, and applying Bayes' Rule, the probability of each object label, c, given the set of observations t, is given by:

$$P(C = c|T_1 = t_1, ..., T_N = t_N) = \alpha\prod_{k=1}^{N} P(T_k = t_k|C = c)P(C = c) \tag{6.1}$$

Where α is a normalizing constant, and $P(C = c)$ can be estimated from the number of times each object label is observed during training. Therefore, touch-only object label prediction is performed by:

$$C_{touch} = \underset{c}{\operatorname{argmax}}\prod_{k=1}^{N} P(T_k = t_k|C = c)P(C = c) \tag{6.2}$$

3.2 Visual model

The visual model uses a bag-of-features approach based on SIFT features [77]. Dictionary learning is performed by applying kmeans to the SIFT descriptors of a set of images which are substantially different to any image in the database. Nearest neighbour is used to turn the set of SIFT descriptors of each image in the database into a histogram of visual 'words', which is also normalised so that

its sum is one. Histograms are used to train an one-vs-all gaussian SVM. During testing, the posterior over labels is predicted applying Platt scaling [92].

Formally, let $s(v)$ be the score given by the SVM corresponding to label c to the visual histogram v of an object's image. Then the probability of label c is estimated as:

$$P(C = c|V = v) = \frac{1}{1 + \exp(As(v) + B)} \tag{6.3}$$

Where A and B are two constants estimated by maximising the log likelihood of the training data (for details, see [92]). The predicted label for vision only is therefore:

$$C_{vision} = \underset{c}{\operatorname{argmax}}\, P(C = c|V = v) \tag{6.4}$$

3.3 Bayesian visuo-tactile integration model

The visuo-tactile integration model (see Fig. 6-4 assumes that T_k $(k = 1, ..., N)$, the events of sensing a given tactile vector t at the k^{th} touch, and V, the event of seeing a photo histogram v, are conditionally independent, given a label $C = c$ is known. Using shortened notation for clarity (read $P(X_i)$ as $P(X_i = x_i)$), this is

$$P(V, T_1, ..., T_N|C) = P(V|C) \prod_{k=1}^{N} P(T_k|C) \tag{6.5}$$

Therefore in order to find the object label, c, that maximises $P(C|V, T)$, apply Bayes's rule,

$$\begin{aligned} C_{pvpt} &= \underset{c}{\operatorname{argmax}}\, P(C|V, T_1, T_2, ..., T_N) \\ &= \underset{c}{\operatorname{argmax}}\, \frac{P(V, T_1, T_2, ..., T_N|C)P(C)}{P(V, T_1, T_2, ..., T_N)} \\ &= \underset{c}{\operatorname{argmax}}\, P(V, T_1, T_2, ..., T_N|C)P(C) \end{aligned}$$

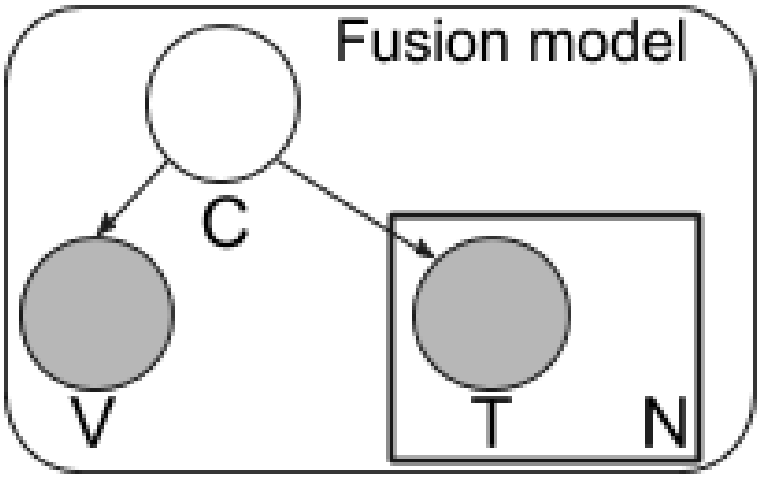


Figure 6-4: *The visuo-tactile integration model. Assume that, given an object label, visual and tactile input marginal likelihoods are conditionally independent. N is the number of tactile images received during testing, i.e. the number of touches.*

And, by using equation (6.5),

$$\begin{aligned} C_{pvpt} &= \underset{c}{\operatorname{argmax}}\, P(V|C)P(C)\prod_{k=1}^{N} P(T_k|C) \\ &= \underset{c}{\operatorname{argmax}}\, P(C|V)P(V)\prod_{k=1}^{N} P(T_k|C) \\ &= \underset{c}{\operatorname{argmax}}\, P(C|V)\prod_{k=1}^{N} P(T_k|C) \end{aligned} \tag{6.6}$$

Where $P(C)$ is estimated by the relative frequency of each label in the training set, $P(T_k|C)$ is obtained from equation (6.4), and $P(C|V)$ is calculated as defined in equation (6.3).

4 The VT-60 database

The visuo-tactile database introduced here consists of 60 household objects, split between 10 classes (See Fig. 6-5).

For each object, photos were taken from 40 viewpoints manually. Tactile images were collected using an open-source inexpensive tactile sensor mounted on a KUKA robotic arm. Each object was automatically explored and touched

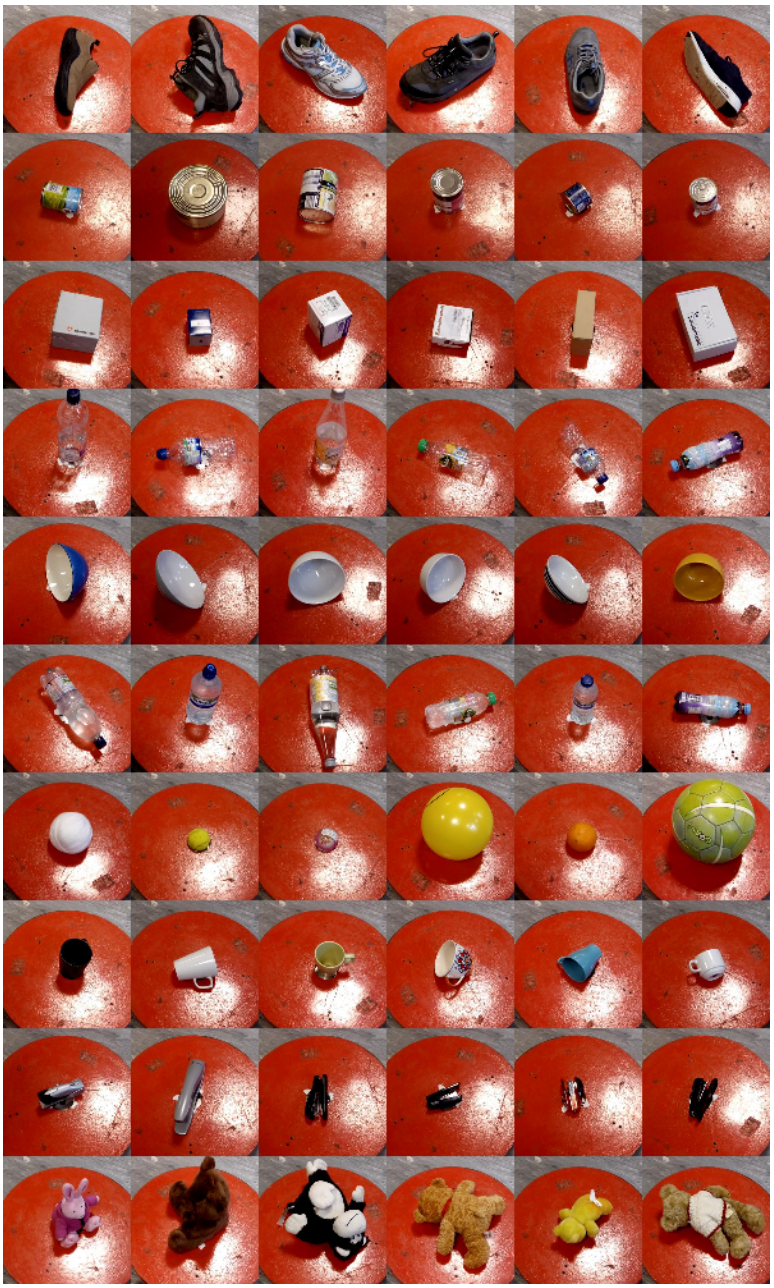


Figure 6-5: *Images of each of the objects in the VT-60 visuo-tactile database, from top to bottom: shoe, can, box, bottle_empty, bowl, bottle_full, ball, mug, stapler, soft_toy.*

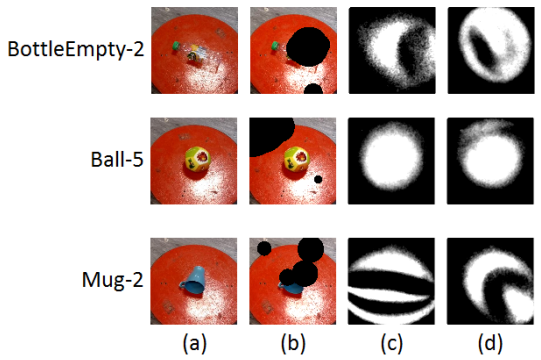


Figure 6-6: *Samples of database objects, showing (a) raw photo, (b) "blotched" photo, (c-d) sample tactile images.*

at 120 randomly chosen points, by means of approaching the object pointing the sensor inwards towards the object's assumed centre (See Fig. 6-3).

When the sensor made contact with the object, a tactile image (the photo of the deformed rubber membrane) was stored, but information about the sensor's location and orientation was discarded. The intention was to obtain a model that would be pose-invariant. Indeed, the orientation of the object was altered periodically during data collection and sometimes even affected by the robot itself. Examples of some objects' photos, "blotched" photos[1] and sample tactile images are shown in Fig. 6-6.

4.1 Experiment 1: Class Recognition

In the first experiment, the aim was to correctly classify an unseen/untouched object. 50 objects (5 instances for each class) were used for training and the remaining 10 (one instance per class) for testing. During training, 40 photos and 60 tactile images of the 50 objects were used. Thus, the prior probability of each object class, $P(C)$, is uniform and set to $\frac{1}{10}$.

During testing, a single photo and up to 30 tactile images of the test object were used to compute the posterior probabilities for each class, according to equation (6.6) and the class label with highest probability was chosen. The disparity between the number of photos and the number of tactile images is due to the fact that photos display the complete object, whilst tactile images correspond to the tactile sensation of just a small portion of the object.

This data split (50/10) was repeated so all objects were tested using 10 dif-

[1]photos where 20% of the pixels were obscured by randomly placed disks

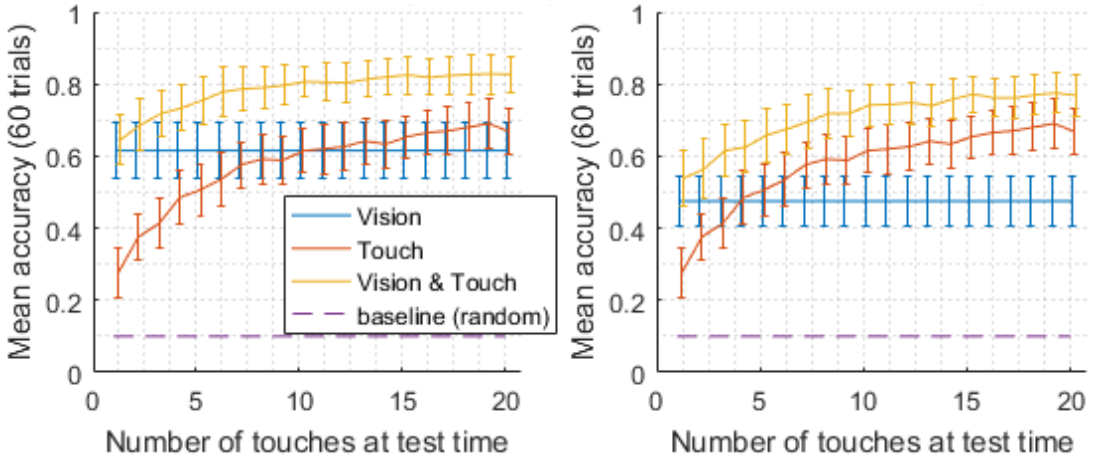


Figure 6-7: *Class recognition (unknown objects) accuracy over 60 trials, as more touches are used at test time, using complete photos (left) and "blotched" photos (right). The bars represent one standard deviation. 54 Tactile images and 9 photos used during training. A single photo and up to 20 tactile images (touches) used during testing. The fusion model outperforms both modalities in all cases.*

ferent photographic viewpoints (for a total of 60 trials). The reported accuracy is the mean proportion of correct class predictions over these 60 trials (baseline random accuracy $\frac{1}{10}$). Fig 6-7 shows mean accuracy for class recognition, as more and more touches are used at test time. Both in the case of unaltered and blotched images, the sensor fusion model outperforms each modality alone significantly. Fig 6-8 shows the confusion matrix for each one of the 60 objects' predicted class, demonstrating clear gains by using the fusion model.

4.2 Experiment 2: Instance Recognition

In this experiment, the aim was to correctly label a previously seen/touched object, considering there are very similar objects in the database. During training, 36 of the 40 photos and 54 tactile images of each object were used. Thus, the prior probability for each object label, $P(C)$, was assumed uniform and set to $\frac{1}{60}$. During testing, 1 of the remaining 4 photos and between 1 and 20 of the remaining 36 tactile images of each object were used to compute the posterior label probabilities, according to equation (6.6). 60 such randomised training/testing splits were run. The reported accuracy is the mean correct object instance predictions over these 60 trials (baseline random accuracy $\frac{1}{60}$). Fig. 6-9 shows accuracy for class recognition, as more and more touches are used at test time. Both in the case of unaltered and blotched images, the sensor fusion model outperforms

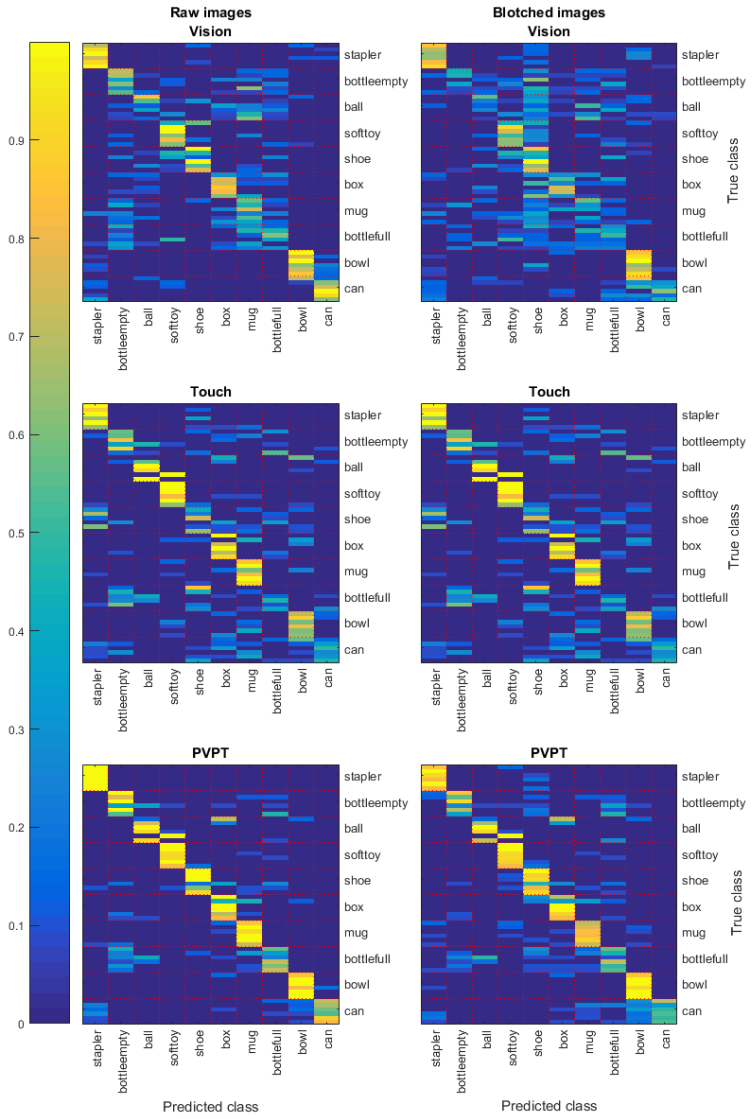


Figure 6-8: *Confusion matrix for class recognition of unseen objects. Each row represents the true identity of the object (grouped by class, 6 of each) and each column represents the predicted class for it. Using Vision-only (top), Touch-only (middle), and the fusion model (PVPT, bottom). Comparison between unaltered photos (left) and "blotched" photos (right). Individual object names removed for clarity. In both contexts, the fusion model reduces uncertainty for most classes.*

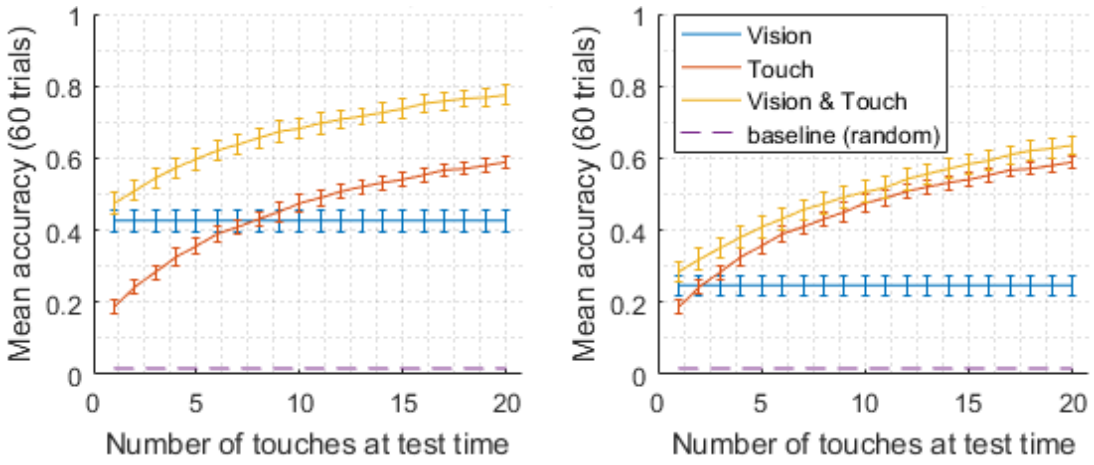


Figure 6-9: *Instance recognition (known objects) accuracy over 60 trials, as more touches are used at test time, using complete photos (left) and "blotched" photos (right). The bars represent one standard deviation. 54 Tactile images and 9 photos used during training. A single photo and up to 20 tactile images (touches) used during testing. The sensor fusion model outperforms vision and touch.*

each modality alone significantly. The confusion matrices for each one of the 60 objects' predicted label are shown in Fig. 6-10 (unaltered photos) and Fig. 6-11 (blotched photos). In all cases, most of the confusion happens within-class and between full and empty bottles.

5 Real-life Application: Underwater Object Class Recognition

Underwater object class recognition is a topic of interest for the law enforcement and defence departments. In order to assess the potential of our system in real-life scenarios, this experiment aims to classify unseen objects (similarly to experiment 1) when the object in question is submerged in murky water (See Fig. 6-12). Training was performed using the "dry" database (VT-60, described above). Testing was performed on 10 unseen objects (one from each class), which were submerged in a tank of water mixed with soil. 30 photos and 30 tactile readings were collected manually for each using a waterproof camera. Each testing trial was performed on a randomly chosen photo and a subset of up to 10 randomly chosen tactile images for each one of these 10 objects. Fig. 6-13 shows the mean accuracy (proportion of correctly classified objects), as more touches are used at test time. The vision accuracy is notably lower than in the "dry"

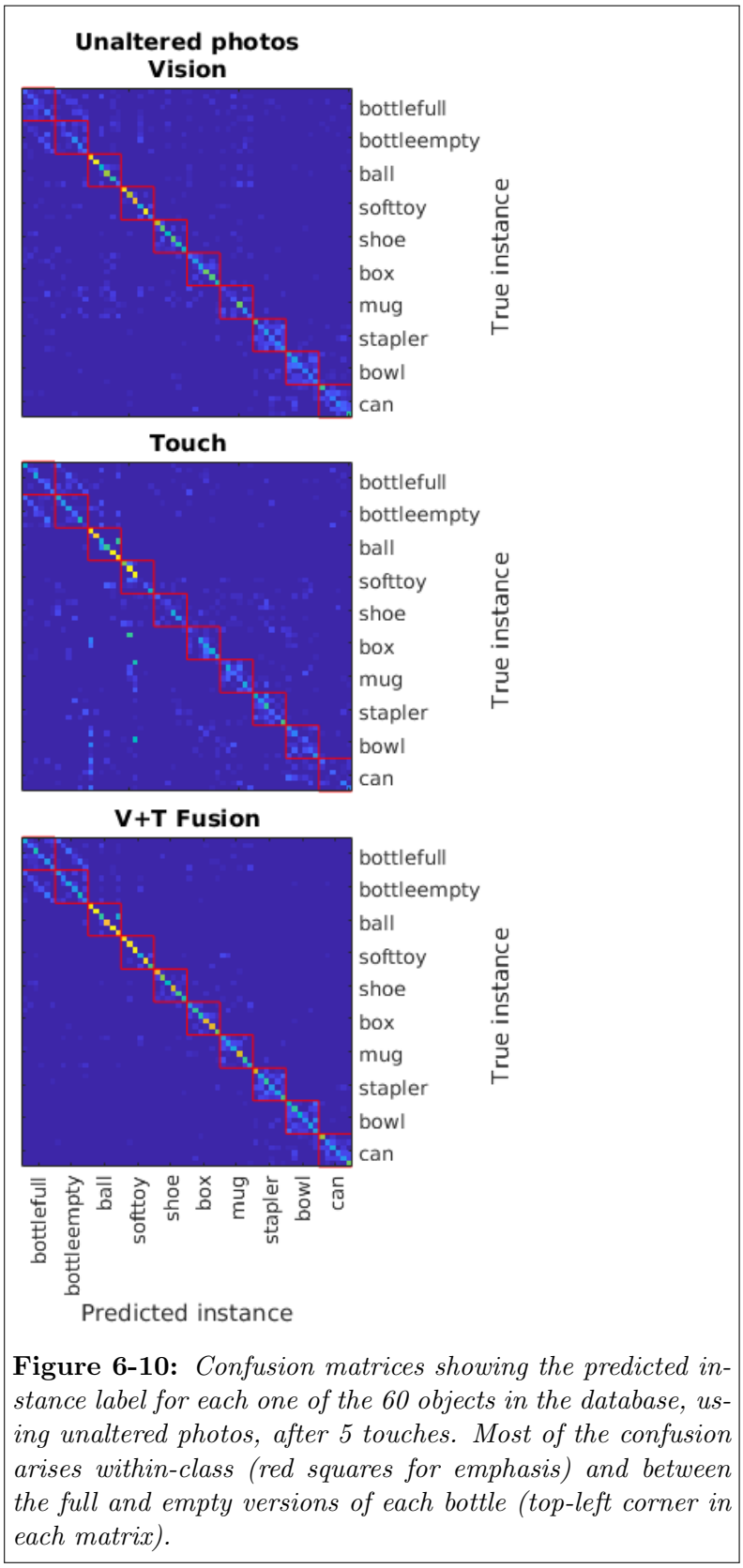


Figure 6-10: *Confusion matrices showing the predicted instance label for each one of the 60 objects in the database, using unaltered photos, after 5 touches. Most of the confusion arises within-class (red squares for emphasis) and between the full and empty versions of each bottle (top-left corner in each matrix).*

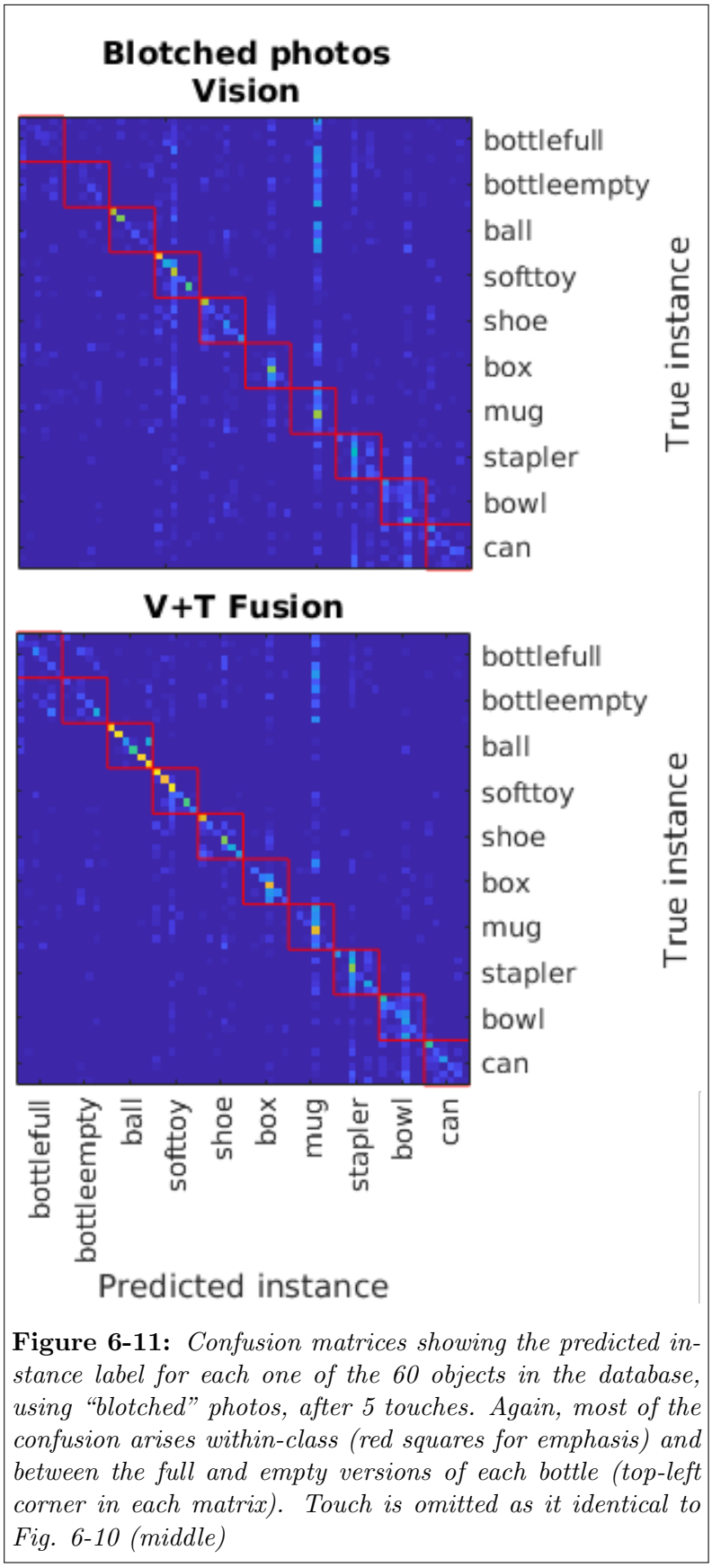


Figure 6-11: *Confusion matrices showing the predicted instance label for each one of the 60 objects in the database, using "blotched" photos, after 5 touches. Again, most of the confusion arises within-class (red squares for emphasis) and between the full and empty versions of each bottle (top-left corner in each matrix). Touch is omitted as it identical to Fig. 6-10 (middle)*

experiment, both using unaltered or "blotched" photos. Tactile classification outperforms vision after 4 touches. The fusion model achieves higher accuracy than either modality alone initially. After 7 touches, there is no significant improvements using the fusion model. Fig. 6-14 shows the confusion matrices after 5 touches, detailing how most of the confusion in tactile classification seems to arise from a tendency towards predicting the soft toys class.

6 Discussion and Conclusions

In the first experiment (classification), both vision and touch struggle to differentiate between empty and full bottles (the content is clear water), in fact full bottles are the poorest performing class. The remarkable misclassification is that pertaining to one of the balls (ball_multicolouredanimal), which is consistently classified as "soft toy" by touch, and presents ambiguity between "soft toy" and "ball" for vision. The pliability of this ball is more on par of that of soft toys in the database, whilst all other balls are much more stiff; furthermore, its multicoloured surface and imperfect roundness may be factors confusing the vision system. In fact it is arguably a more accurate classification to label it as a soft toy. The vision system, in general, performs poorly classifying balls, perhaps due to their uniform images resulting in few SIFT descriptors. This confusion is usually resolved by touch, with the exception of the two largest balls, which are sometimes misclassified as boxes, which may be linked to their low curvature, becoming indistinguishable from a flat surface.

One noticeable exception is ball_yellowbeachball, which vision classifies confidently as a ball, yet touch classifies it as a box (perhaps due to its large size, it may appear similar to the flatness of boxes), to the point that it confuses the fusion model sufficiently away from the correct class label.

The second experiment demonstrates the ability of the system to perform traditional object recognition for a large database, even when several objects are similar (belong to the same class). The confusion matrices in Fig. 6-10 and Fig. 6-11 show that confusion mostly arises within-class and between bottles (there is almost no distinction between empty and full bottles). In most cases, the sensor fusion reduces uncertainty, as evidenced also by the mean accuracy, shown in Fig. 6-9.

The underwater experiment is a demonstration of the potential practical applications of this approach, attempting to classify unseen objects. Vision is here ac-

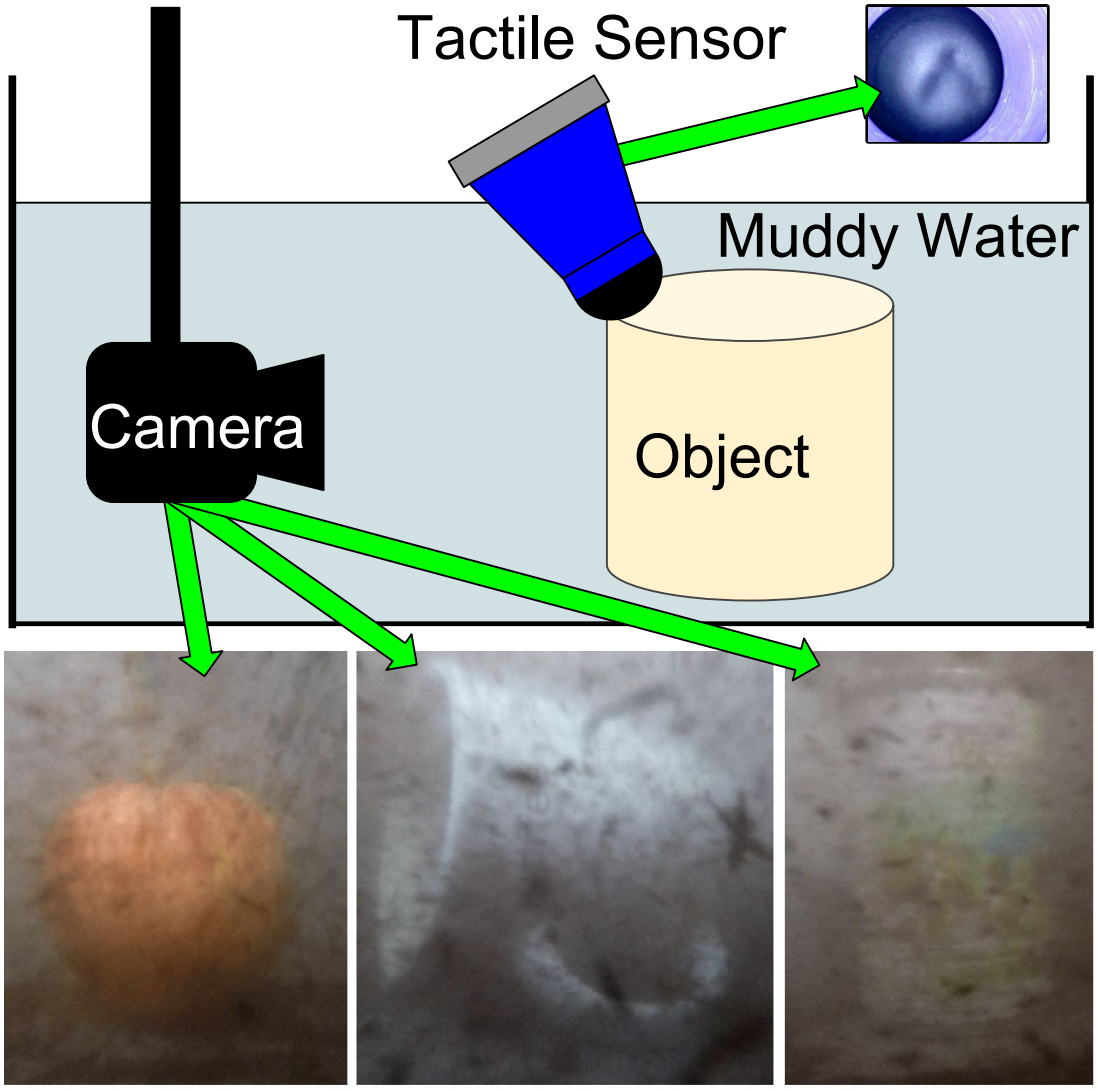


Figure 6-12: *Photos and tactile reading procedure diagram of an object submerged in muddy water. Both tactile readings and photos were collected manually.*

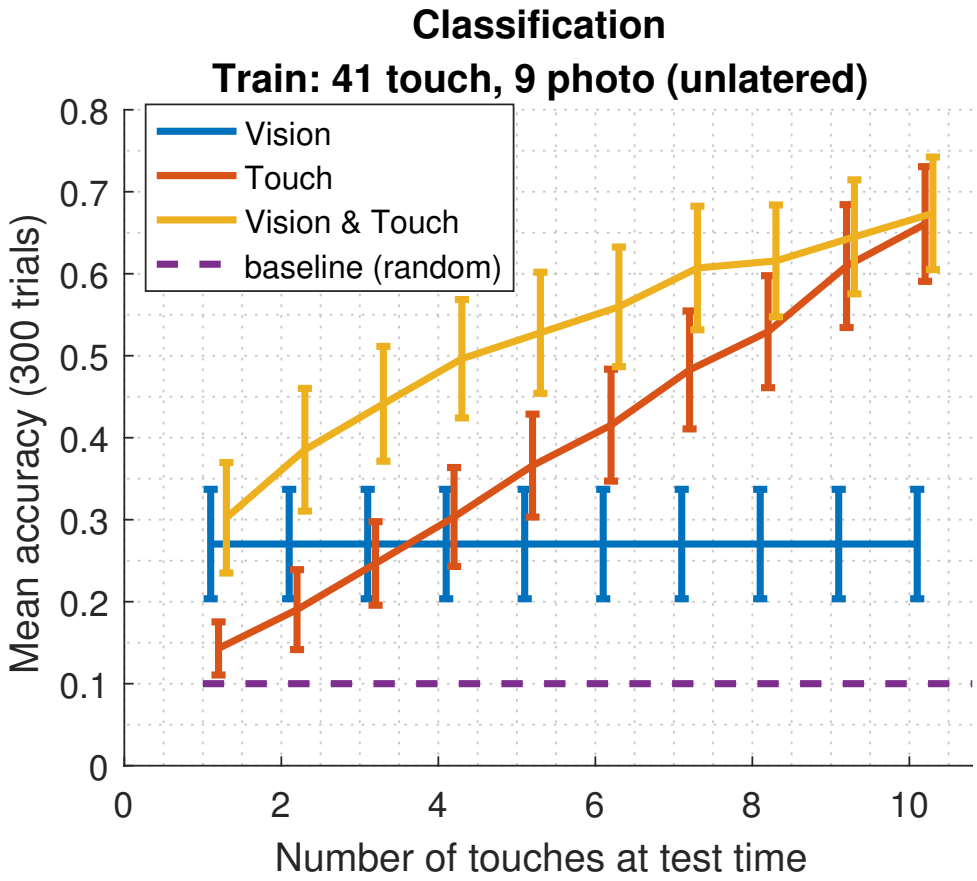


Figure 6-13: *Mean accuracy of classification for objects submerged in muddy water, compared to the number of touches used at test time, over 300 trials (distinct photo+touches combinations). The bars represent one standard deviation. Sensor fusion provides higher accuracy in most cases.*

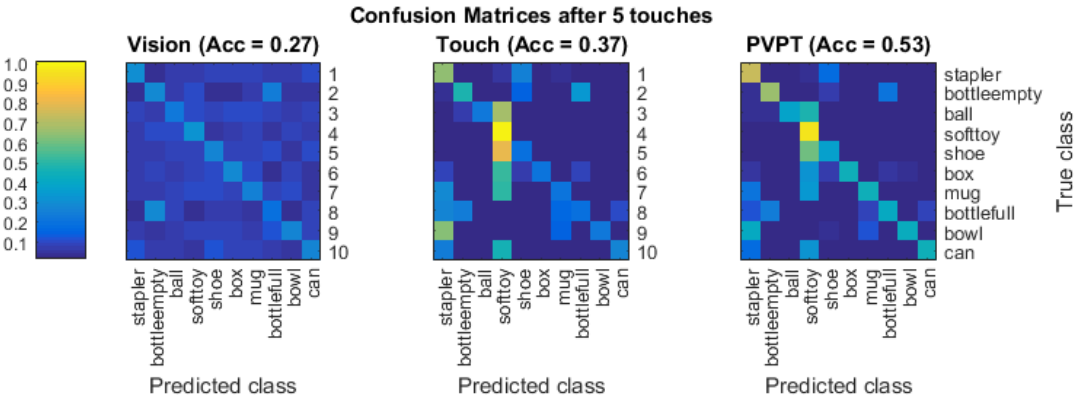


Figure 6-14: *Confusion matrix for objects submerged in muddy water, after 5 touches used at test time, over 300 trials (distinct photo+touches combinations).*

tually impaired by the presence of water, debris and mud, which results in a very low base visual recognition accuracy (0.27). Tactile accuracy is also marginally lower than in the "dry" experiment. The confusion matrix (Fig. 6-14) shows a tendency to classify objects as "soft toys" as a main source of uncertainty, this may be due to the irregular pressure applied during data collection (since it was manual). Even under these unfavourable conditions, at all points, the multi-modal fusion system outperforms individual modalities.

In all experiments, distinction between empty bottles and those full of water proves particularly difficult (in the underwater experiment, no credible discernment was achieved by either vision, touch or the fusion model).

Visuo-tactile object class and instance recognition were shown to achieve high accuracy using an inexpensive tactile sensor and a simple Bayesian sensor fusion model. This is the first time class recognition is attempted using individual touches and visuo-tactile fusion. The model demonstrates the advantages of multi-modal object representation in both contexts. The underwater experiment results should be considered as preliminary, as only one object per class was tested. It serves as a proof-of-concept, highlighting the potential of the approach, and it warrants further research. The database is made available with this paper so further attempts may improve on these results.

Future work will compare this probabilistic model to discriminative and generative neural models (such as deep learning). There is scope to improve the underwater experiment by extending the model to consider the context as a factor to be identified, or perhaps by using domain adaptation [9].

6.5 Further Results: visuo-tactile object classification with deep-learning computer vision

The previous section established the gains to be made by using visuo-tactile sensor fusion in terms of accuracy of recognition and classification. The visual model employed was sufficient for the specific context and question, but was by no means a state-of-the-art vision system. Therefore the question arises, is the visuo-tactile fusion model beneficial even for the best vision models available today?

To date, deep learning models consistently achieve the highest scores in various well-known competitions (e.g. [50]). Of those publicly available, the choice made here was to use VGG16 [76] as it achieves very high scores in image classification, it is simple enough to be usable in consumer-grade GPUs (NVIDIA GTX 765M), and is readily deployable with the Keras high-level interface libraries [19]. Training the complete net (from random initial weights) would require vast amounts of computing power, memory and data. A simpler approach is to fix most layers of the net (thus using them as a feature extractor). Two alternatives are then possible: either fine-tune (initialise to the pre-trained weights and allow for small changes) the topmost N layers with new data or to train a classifier net on top of the last convolutional layer. The first option is suitable for fine-grained recognition (e.g. learning to distinguish between subspecies of fish) and requires a larger amount of data than it is available in this context. The aim here is not to perform fine-grained recognition but to deploy a classifier that can rapidly learn to distinguish between fairly distinct classes from few samples. The second option is therefore chosen.

The version of VGG used was obtained already pre-trained on the Imagenet data set [100], as available in the Keras software library [19]. All except the top fully connected layers of the model were kept and fixed, and three dense layers were stacked on top: 256, 32, and 10 units respectively, all fully connected, using regularised linear activation for the first two, and softmax activation for the final layer (See Fig. 6-15). Training was performed using the Adadelta optimiser [131].

For the purposes of sensor fusion, the net's final layer (a softmax layer) units' responses were interpreted as the posterior probability for each class. Training and testing were performed using the same procedure as described in the previous sections: multiple 50/10 object data splits, and computing the average accuracy of classification.

Layer (type)	Output Shape	Param #	Trainable
input_1 (InputLayer)	(None, 224, 224, 3)	0	No
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792	No
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928	No
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0	No
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856	No
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584	No
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0	No
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168	No
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080	No
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080	No
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0	No
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160	No
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808	No
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808	No
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0	No
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808	No
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808	No
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808	No
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0	No
flatten (Flatten)	(None, 25088)	0	No
fc1 (Dense)	(None, 256)	6422784	Yes
do1 (Dropout, 0.25)	(None, 256)	0	Yes
fc2 (Dropout, 0.25)	(None, 32)	8224	Yes
do2 (Dropout, 0.25)	(None, 32)	0	Yes
predictions (Dense)	(None, 10)	330	Yes

Total params: 21,146,026
Trainable params: 6,431,338
Non-trainable params: 14,714,688

Figure 6-15: *Architecture of the deep net used for vision classification. The base net used was VGG16, pretrained on Imagenet. The last 3 layers of VGG-16 were removed and replaced by custom layers ('fc1', 'fc2', and 'predictions').*

Unaltered	No touches	1 touch	3 touches	5 touches	10 touches	15 touches	20 touches
Touch-only	-	0.27	0.41	0.48	0.59	0.65	0.68
BoW	0.62	0.65	0.73	0.76	0.81	0.84	0.84
VGG16-40	0.95	0.96	0.95	0.95	0.95	0.95	0.96
VGG16-20	0.88	0.87	0.88	0.89	0.91	0.91	0.91
VGG16-10	0.85	0.84	0.85	0.86	0.88	0.89	0.89
VGG16-5	0.71	0.76	0.77	0.78	0.79	0.81	0.82

Table 6.1: *A comparison of mean accuracies between Touch-only, pure vision using the Bag-of-Words model presented in Chapter 5 (Vision BoW), and pure vision using a fine-tuned deep neural net (VGG16-xx), where the suffix indicates the number of photos used during training, starting with 40 (VGG16-40), the same as the BoW model, and down to 5 (VGG16-5). Columns 2-7 represent the accuracies using the sensor fusion model presented in this chapter, with more and more touches allowed during test time.*

Blotched	No touches	1 touch	3 touches	5 touches	10 touches	15 touches	20 touches
Touch-only	-	0.27	0.41	0.48	0.59	0.65	0.68
BoW	0.47	0.55	0.63	0.67	0.75	0.77	0.77
VGG16-40	0.69	0.72	0.74	0.76	0.78	0.80	0.81
VGG16-20	0.52	0.57	0.61	0.65	0.70	0.74	0.76
VGG16-10	0.45	0.48	0.55	0.58	0.63	0.66	0.66
VGG16-5	0.40	0.48	0.55	0.59	0.66	0.71	0.73

Table 6.2: *A comparison of mean accuracies between Touch-only, pure vision using the Bag-of-Words model presented in Chapter 5 (BoW), and pure vision using a fine-tuned deep neural net (VGG16-xx), where the suffix indicates the number of photos used during training, starting with 40 (VGG16-40), the same as the BoW model, and down to 5 (VGG16-5). Columns 2-7 represent the accuracies using the sensor fusion model presented in this chapter, with more and more touches allowed during test time.*

Allowing all 40 photos of each object to be used for training resulted in almost-perfect accuracy, and was not informative. Therefore, the same experiment was run, reducing the number of photos used during training. The resulting accuracies are shown in Table 6.1 for unaltered photos and Table 6.2 for blotched photos.

In the case of unaltered photos (Table 6.1), the deep vision based model outperforms the basic visual model used in the previous chapters (Vision BoW) easily even with only 5 photos for training, obtaining an accuracy of 0.71 (bottom row). In the case of blotched photos (Table 6.2), the deep vision model requires 20 photos at training to beat the BoW model. The deep vision model is performing noticeably well, as expected.

In terms of the fusion model as applied to the deep vision model, in most cases, it provides gains. For unaltered photos (Table 6.1), all versions of the vision model are further improved by the fusion model. The most improved was

VGG16-5, which presents an error reduction of between 16% and 38% (1 to 20 touches). The notable exception is VGG16-40, whose accuracies are already so high (0.95) that improvement is difficult. In fact, it is significant that no accuracy drops are recorded, as there could have been a 'confusion' effect if the touch model was performing particularly poorly. This may be due to the fact that deep nets are trained with categorical 'one-hot' encoding for the output leading to high certainty for clear classifications. In fact, over 95% of activations of the output layer for VGG16-40 using unaltered images for correct classifications are within 10^{-4} of 1, suggesting that the net is producing high certainty predictions (close to 1 probability for the predicted class and close to 0 for all others). Furthermore, approximately two-thirds of the activations of incorrect classifications are also virtually 1, making it impossible for the posterior product fusion model to correct the final prediction. The deep net is overconfident in its predictions.

For blotched photos (Table 6.2), the deep vision model does not perform so well to begin with (0.69 using 40 training photos), and here the improvements of the fusion model are much more marked, reducing error by 13-55% (1 to 20 touches) for VGG16-5, and by 10-39% (1 to 20 touches) for VGG16-40.

To conclude, the fusion model is again providing clear gains in accuracy, even for a powerful deep-learning vision system, and these are most notable when images are "blotched".

Chapter 7

Discussion and conclusions

7.1 Hypothesis testing and contributions

Recall that the hypotheses of this thesis (see Section 1.1) read:

1. Non-grasping tactile object classification is feasible with a simple, low cost tactile sensor.

2. A simple probabilistic graphical model for the integration of tactile and visual robotic perception is likely to yield higher accuracy object instance recognition and object classification than either modality alone.

In attempting to address hypothesis 1, a number of experiments were conducted using a novel, open-source, low-cost, optics-based tactile sensor. Using individual touching contacts (non-grasping), data were collected initially about tactile shapes (flat, edge, etc.) and then about objects. For the first time in the literature, such an approach produced significant object classification results (correctly recognising a new mug, for example, when said mug was not in the training set, while other mugs were).

In attempting to address hypothesis 2, three competing fusion models were compared: a basic heuristic, a model from the literature and a proposed model based on a probabilistic approach. Initially (Chapter 5) object recognition was attempted using a small data set, on which the superiority of the proposed probabilistic model was established. During this stage, all three models produced higher accuracy when combining modalities than using each modality alone. The chosen probabilistic model was then tested on a larger, more challenging data set both for instance recognition and classification. Once again, accuracy of recog-

nition was higher for the fusion model than for either modality alone, even in a test scenario involving objects submerged in dirty waters. Similar results were obtained when the simple vision model was replaced with an advanced deep-learning vision model, adapted to produce an output that can be used as a probability distribution over classes.

As a summary, the contributions reported in this thesis are:

- the design and construction of a novel, low-cost tactile sensor based on the image of the shading pattern of the deformation of a rubber membrane,
- the invention of a tactile object recognition system based on Zernike features using said sensor,
- the publication of the largest visuo-tactile household object database to date,
- the first example of tactile and visuo-tactile object classification, using a simple fusion model based on the assumption of conditional independence of sensory data, given a label,
- the demonstration of potential for underwater object classification, and
- the demonstration of the fusion model using deep learning based machine vision.

7.2 Discussion

Technological challenges regarding the replication of the TacTip sensor and the failure to reproduce the fine papillae in this attempt resulted in the creation of a simpler, cheaper, tactile sensor, which was named the "BathTip". The sensor's simplicity was not a barrier to its potential, as was demonstrated in the various experiments outlined in this thesis. Its capability for simple shape recognition was demonstrated in Chapter 3. The basic tactile sensations proposed (nothing, flat, edge, flat-to-edge, corner, cylinder) are arbitrary, but the clear clustering demonstrated and the high classification scores achieved showed the sensor had potential for shape discrimination on a par (or surpassing) the TacTip itself. This was sufficient evidence to adopt the use of Zernike-PCA features for the subsequent stages of the project.

The tactile object recognition model, introduced in Chapter 4, builds on the initial findings by using Zernike features and PCA to transform each tactile image into a vector of 20 numbers. The representation of an object is considerably more complex than the representation of shapes, as it must combine multiple such vectors at once. The solution proposed was using a likelihood function based on a sum of Gaussians, a simple Bayesian approach and Maximum-a-Posteriori, to recognise objects. The paper reported the highest-to-date non-grasping tactile accuracy for recognition for a small set of household objects. Given the disparity between experimental setups, it cannot be claimed that the system is generally superior to other works compared, as the objects are different. This is a common problem in the field of tactile recognition, since data collection is expensive and the data formats are intrinsically linked to the sensors used, so data collection cannot be crowd-sourced. It is not uncommon to see data sets of between 5 and 20 objects, orders of magnitude smaller than their counterparts in machine vision. This thesis is a first effort in providing larger data sets to the community, but other attempts are ongoing [14], and a large dataset for fabrics has been recently been made available [**?**].

The tactile model was then extended to incorporate vision, providing a full sensor fusion system. The vision model chosen combined histograms of SIFT/SURF features with an SVM (implementing a pipeline similar to [28]), which is simple and not claiming to be state-of-the-art. Indeed the data sets used are too small for the world of machine vision and would be overwhelmed by powerful vision systems. On occasions, the simple vision system deployed needed to be impaired by introducing artificial "blotches" (occlusion of parts of the image by using black blobs, emulating visual impairment) so it would not dominate. The fusion model was based on the assumption that the visual and tactile data observations are conditionally independent, given an object identity/class. It was then shown that, in practice, in order to predict an object identity/class, it suffices to maximise a product of modality posterior probabilities (Chapter 5, or a product of the visual posterior and the tactile likelihood (6). A comparison was drawn between this proposed method and two others: a heuristic method which simply calculated a weighted average of the probabilities of each object for vision and touch, and a concatenation method combined with a weighted nearest neighbour classifier (similar to the work in [127]). The proposed model obtained the best highest accuracy. The paper validated the choice of the proposed probabilistic model, warranting its continued use in the next stage of the project.

The tactile model was then shown (Chapter 6) to be able to scale to larger databases and was the first example of tactile classification reported in the literature, and the fusion model was the first example of visuo-tactile classification. The confusion matrices shown give a clear idea that even for the most challenging of problems (such as distinguishing between empty and full bottles), there were accuracy gains made by the multi-modal approach. The only scenario where the fusion model did not provide a significant improvement was when attempting recognition (i.e. correctly identifying one in 60 objects) with impaired vision (i.e. "blotched images"). Since the blotches are placed at random locations and are required to cover 20% of the pixels, it may very well be impossible in some cases to discern the object at all. Vision performs so poorly in some cases, that tactile input is likely the only true discerning factor, so their combination is not superior.

A source of error may be the overconfidence of the individual modalities, i.e. predicting erroneous labels with high probability and giving vanishingly low probabilities to the correct label. Recall that class prediction can be performed by maximising the product of the posterior probabilities for vision and touch (see Chapter 5). Therefore, probabilities close to zero are problematic as the product will tend to vanish. Fig. 7-1 shows three key examples of predictions from each modality, using unaltered photos, that may illustrate this point.

In the case of the empty bottle, neither vision nor touch produce an accurate prediction, mostly due to the confusion with the similar class 'bottle empty'. However, the fusion model predicts the class correctly. Notice that in this example, touch is producing something close to a uniform posterior (low certainty), while vision is mostly undecided as to whether it is an empty bottle or a full bottle. In the case of the orange ('ball'), the high confidence produced by the vision model is insufficient to lead to a correct result in the fusion prediction. The touch model is clearly overconfident in the subset of erroneous predictions ('box', 'bowl', 'can'). However, vision is partly to blame, since it is assigning a non-trivial probability to the 'bowl' class. In the case of the teddy bear ('soft toy'), the tactile system has the correct prediction (even after only one contact), but high uncertainty. Therefore, the overconfidence of the vision system (a combination of a high probability for 'bottle full' and a low probability for 'soft toy') results in an inaccurate prediction overall.

Considering potential applications of the multi-modal classification approach, and following the success of recent work in underwater tactile recognition of objects given their 3D models [1], the system was tested and shown capable of

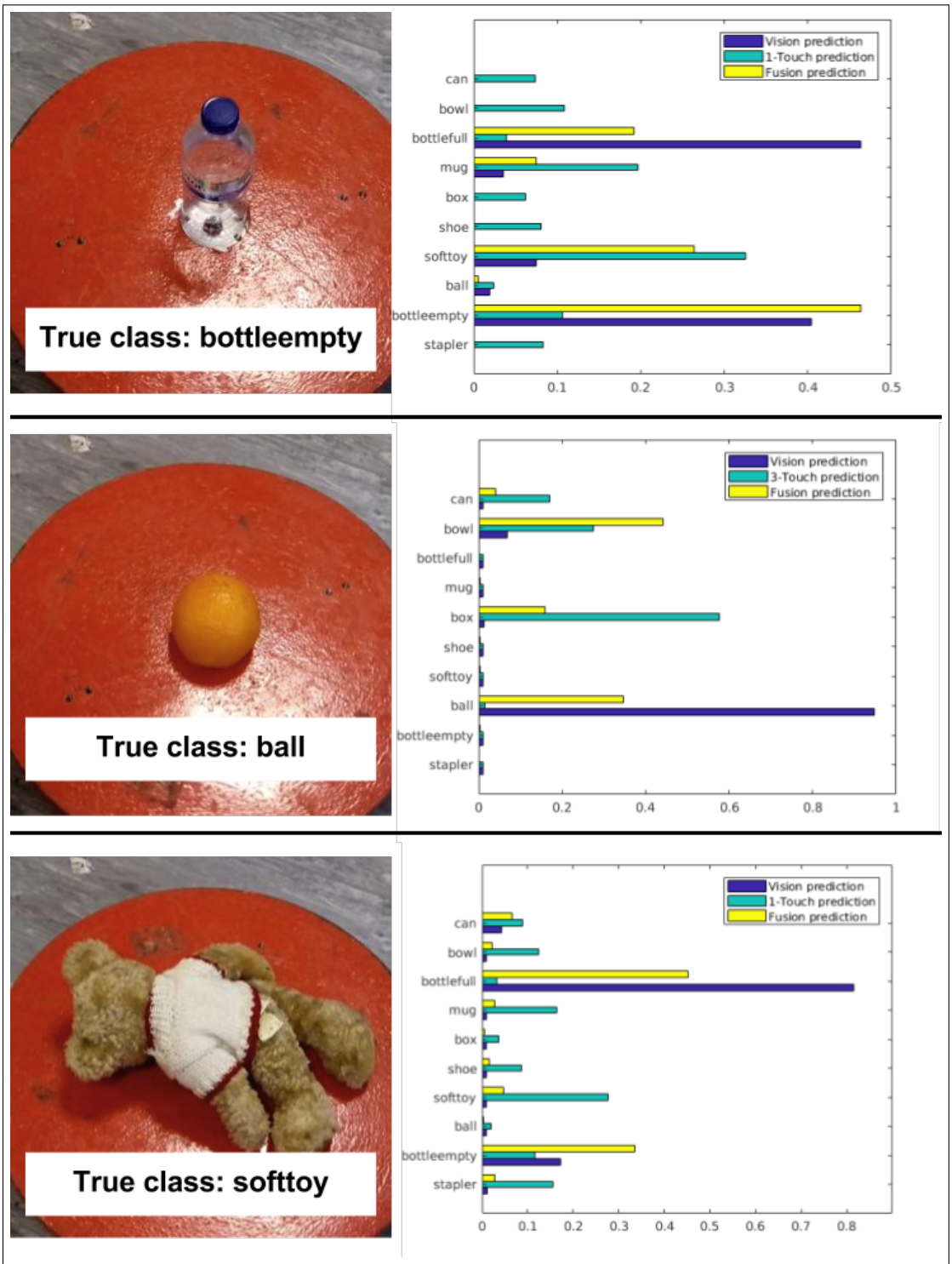


Figure 7-1: *Three significant examples of class predictions for objects in the VT-60 database. An empty bottle (top) is correctly classified by the fusion model, despite incorrect labels from both modalities. An orange (middle) is incorrectly classified by the fusion model, despite it being correctly classified by vision. A soft toy (bottom) is incorrectly classified by the fusion model, despite being correctly classified by touch.*

classifying 10 new objects which were submerged in murky water, providing a proof-of-concept for a potential practical application of the technology. However, as the data set was small, this should be considered preliminary only.

Further tests conducted on the data set using an adapted deep neural net (VGG16) showed (see Section 6.5) the sensor fusion model can provide gains in accuracy even for these advanced technologies. These gains were most marked for the worst performing versions of the deep net, possibly highlighting some limitations of interpreting soft-max layers as probability distributions, if one-hot encoding is used for training, as the neural net was often overconfident in its predictions, even for erroneous classifications. Various alternative architectures are available for image classification, and a coverage of all their merits goes beyond the scope of this project, the reader is referred to the review by Schmidhuber et al. [103] for further information. As shown in Table 6.1, the chosen architecture was able to classify objects with accuracies of 0.71 with only 5 photos used during training, remarkably superior to the simple visual model used in Chapters 5 and 6. The significant result here is that, even for this advanced machine vision method, the fusion model proposed shows gains in accuracy.

7.3 Conclusion

Experiments reported in Chapters 3-6 show that an inexpensive, simple tactile sensor is indeed capable of performing non-grasping tactile shape recognition, object instance recognition, and object classification in small and medium sized data sets. The fusion model designed surpasses a baseline heuristic model and the only other comparable approach found in the literature [127] for this data set. It also consistently attains a higher accuracy of recognition and classification than either vision or touch alone, as supported by numerous experiments reported in Chapters 5-6. Chapter 6 provides evidence of accuracy gains even when the vision model used is an advanced deep learning net.

In addition to the above, the results discussed in Chapter 5 suggest in certain circumstances (when both vision and touch perform poorly independently), *learning efficiency* is also higher. That is, the number of training samples required to obtain a similar accuracy is, overall, smaller. Chapters 5 and 6 provide evidence to conclude that the sensor fusion model provides the highest gains in accuracy when neither modality dominates.

Overall, visuo-tactile integration is considered to be a promising prospect

for object representation in robotics, in particular with regards to robustness to sensor failure or underperformance. If robots are to operate in a manner that is resilient to these challenges, multi-modal representations are, I conclude, fundamental. It is my hope that this thesis: the database published, the sensor and the model formulations provide a step in this direction.

7.4 Limitations and further work

The strengths outlined above should be put in context, and the limitations of each contribution must be understood to assess the significance of this work and the potential for further research:

- The sensor size is relatively large, making it unsuitable for mounting on humanoid hands, for example. Further work could explore opportunities for miniaturisation, such as was the case for the TacTip [125].

- Details of parts, materials, structure and components is given in full. However testing pertaining sensor drift or robustness to third factors such as lighting and varied pressure profiles was not performed. Further work could explore ways of extracting force vectors from the deformation shading pattern. One possibility would be to aim to reconstruct the 3D shape of the rubber membrane, as was done by Ferrier et al. [37].

- The conclusions reached here are limited to the given context (household objects, BathTip sensor, controlled conditions), further work should focus on extending them to other sensors and conditions, and/or to a general case, whensoever a more unified approach to robotic tactile sensing is reached.

- Scalability to very large data sets: the tactile model requires maintaining in memory a small vector (of dimensionality 20) for each tactile contact made with each known object, amounting to approximately 6Kb per object). Object label inference, as described in Chapter 4, requires an entire pass over all known data samples. This may be prohibitive for large-scale data sets. Since tactile data collection was time-expensive, no attempt was made to optimise the model for efficiency.

- Limited scope of the vision models: the first visual model employed here is relatively simple compared to the large number of approaches in the field.

The conclusions reached should not be interpreted as furthering the field of machine vision, but instead as testing the potential of machine touch and machine visuo-tactile sensing. The deep learning approach considered is one of many possibilities and therefore the conclusions cannot generalise. Further work could explore if gains can also be made over state-of-the-art machine vision approaches in general, including the most powerful neural nets [50], or deformable part models [36]. A completely probabilistic approach would also be of interest, perhaps using Gaussian Processes [96].

- Only one exploratory procedure used. According to Lederman el at. [70], humans use six types of interactions with objects when learning their haptic properties (press, stroke, static contact, enclosure/grasping, weighing, contour following). In this study, only static contact was used. There remains much work to be done with regards to dynamic tactile responses, which should be considered for further work.

References

[1] A. Aggarwal, P. Kampmann, J. Lemburg, and F. Kirchner, *Haptic object recognition in underwater and deep-sea environments*, Journal of Field Robotics, 32 (2015), pp. 167–185.

[2] P. K. Allen, *Integrating vision and touch for object recognition tasks*, International Journal of Robotics Research, 7 (1988), pp. 15–33.

[3] P. K. Allen and R. Bajcsy, *Object recognition using vision and touch*, in IJCAI, 1985.

[4] P. K. Allen, A. T. Miller, P. Y. Oh, and B. S. Leibowitz, *Integration of vision, force and tactile sensing for grasping*, International Journal of Intelligent Machines, 4 (1999), pp. 129–149.

[5] T. Assaf, C. Roke, J. Rossiter, T. Pipe, and C. Melhuish, *Seeing by touch: Evaluation of a soft biologically-inspired artificial fingertip in real-time active touch*, Sensors (Switzerland), 14 (2014), pp. 2561–2577.

[6] H. Barron-Gonzalez and T. Prescott, *Discrimination of social tactile gestures using biomimetic skin*, in Conference Towards Autonomous Robotic Systems, Springer, 2013, pp. 46–48.

[7] Y. Bekiroglu, D. Kragic, and V. Kyrki, *Learning grasp stability based on tactile data and HMMs*, in Proceedings - IEEE International Workshop on Robot and Human Interactive Communication, sep 2010, pp. 132–137.

[8] Y. Bekiroglu, J. Laaksonen, J. A. Jørgensen, V. Kyrki, and D. Kragic, *Assessing grasp stability based on learning and haptic data*, IEEE Transactions on Robotics, 27 (2011), pp. 616–629.

[9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, *A theory of learning from different domains*, Machine Learning, 79 (2010), pp. 151–175.

[10] A. Bierbaum, I. Gubarev, and R. Dillmann, *Robust shape recovery for sparse contact location and normal data from haptic exploration*, in 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, sep 2008, pp. 3200–3205.

[11] A. Bierbaum, K. Welke, D. Burger, T. Asfour, and R. Dillmann, *Haptic exploration for 3D shape reconstruction using five-finger hands*, in Proceedings of the 2007 7th IEEE-RAS International Conference on Humanoid Robots, HUMANOIDS 2007, 2008, pp. 616–621.

[12] M. Bjorkman, Y. Bekiroglu, V. Hogman, and D. Kragic, *Enhancing visual perception of shape through tactile glances*, in IEEE International Conference on Intelligent Robots and Systems, 2013, pp. 3180–3186.

[13] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-PLUS Illustrations*, 1997.

[14] A. Burka, S. Hu, S. Helgeson, S. Krishnan, Y. Gao, L. A. Hendricks, T. Darrell, and K. J. Kuchenbecker, *Design and Implementation of a Visuo-Haptic Data Acquisition System for Robotic Learning of Surface Properties*, Iee, (2016), pp. 1–3.

[15] M. C. Burl, M. Weber, and P. Perona, *A probabilistic approach to object recognition using local photometry and global geometry*, Jet Propulsion, (1998).

[16] S. Caselli, C. Magnanini, and F. Zanichelli, *Haptic Object Recognition with a Dextrous Hand Based on Volumetric Shape Representations*, in IEEE Int. Conf. on Multisensor Fusion and Integration, Las Vegas, NV, 1994, pp. 2–5.

[17] Y. Cheng, C. Su, Y. Jia, and N. Xi, *Data correlation approach for slippage detection in robotic manipulations using tactile sensor array*, in IEEE International Conference on Intelligent Robots and Systems, vol. 2015-Decem, 2015, pp. 2717–2722.

[18] S. Chitta, M. Piccoli, and J. Sturm, *Tactile object class and internal state recognition for mobile manipulation*, in Proceedings - IEEE International Conference on Robotics and Automation, may 2010, pp. 2342–2348.

[19] F. Chollet, *Keras*, 2015.

[20] C. Chorley, C. Melhuish, T. Pipe, and J. Rossiter, *Development of a Tactile Sensor Based on Biologically Inspired Edge Encoding*, Design, (2009), pp. 1–6.

[21] C. Chorley, C. Melhuish, T. Pipe, and J. Rossiter, *Tactile edge detection*, in Proceedings of IEEE Sensors, 2010, pp. 2593–2598.

[22] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, N. Fitter, J. C. Nappo, T. Darrell, and K. J. Kuchenbecker, *Using robotic exploratory procedures to learn the meaning of haptic adjectives*, in Proceedings - IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 2013, pp. 3048–3055.

[23] F. B. Colavita, *Human Sensory Dominance*, Perception & Psychophysics, 16 (1974), pp. 409–412.

[24] T. Corradi, P. Hall, and P. Iravani, *Tactile features: Recognising touch sensations with a novel and inexpensive tactile sensor*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8717 LNAI, Springer Verlag, 2014, pp. 163–172.

[25] T. Corradi, P. Hall, and P. Iravani, *Bayesian tactile object recognition: Learning and recognising objects using a new inexpensive tactile sensor*, in 2015 IEEE International Conference on Robotics and Automation (ICRA), vol. 2015-June, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 3909–3914.

[26] T. Corradi, P. Hall, and P. Iravani, *Object recognition combining vision and touch*, Robotics and Biomimetics, 4 (2017), p. 2.

[27] L. Cramphorn, B. Ward-Cherrier, and N. F. Lepora, *Tactile manipulation with biomimetic active touch*, in 2016 IEEE International Conference on Robotics and Automation (ICRA), May 2016, pp. 123–129.

[28] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, *Visual categorization with bags of keypoints*, Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision, (2004), pp. 59–74.

[29] R. S. Dahiya, P. Mittendorfer, M. Valle, G. Cheng, and V. J. Lumelsky, *Directions toward effective utilization of tactile skin: A review*, 2013.

[30] S. Decherchi, P. Gastaldo, R. S. Dahiya, M. Valle, and R. Zunino, *Tactile-Data Classification of Contact Materials Using Computational Intelligence*, IEEE Transactions on Robotics, 27 (2011), pp. 635–639.

[31] A. Drimus, G. Kootstra, A. Bilberg, and D. Kragic, *Design of a flexible tactile sensor for classification of rigid and deformable objects*, Robotics and Autonomous Systems, 62 (2014), pp. 3–15.

[32] M. O. Ernst and M. S. Banks, *Humans integrate visual and haptic information in a statistically optimal fashion.*, Nature, 415 (2002), pp. 429–433.

[33] Fei-Fei Li and P. Perona, *A Bayesian Hierarchical Model for Learning Natural Scene Categories*, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, IEEE, 2005, pp. 524–531.

[34] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, *Object Detection with Discriminatively Trained Part Based Models*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32 (2009), pp. 1–20.

[35] R. Fergus, *Classical Methods for Object Recognition*, in ICCV 2009 Short Course on Recognizing and Learning Object Categories, Kyoto, 2009.

[36] R. Fergus, P. Perona, and A. Zisserman, *Object class recognition by unsupervised scale-invariant learning*, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, jun 2003, pp. 264–271.

[37] N. J. Ferrier and R. W. Brockett, *Reconstructing the Shape of a Deformable Membrane from Image Data*, The International Journal of Robotics Research, 19 (2000), pp. 795–816.

[38] M. Fischler and R. Elschlager, *The Representation and Matching of Pictorial Structures*, IEEE Transactions on Computers, C-22 (1973), pp. 67–92.

[39] M. a. Fischler and R. C. Bolles, *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*, Communications of the ACM, 24 (1981), pp. 381–395.

[40] K. Fukushima, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, Biological Cybernetics, 36 (1980), pp. 193–202.

[41] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, *Deep learning for tactile understanding from visual and haptic data*, in Proceedings - IEEE International Conference on Robotics and Automation, vol. 2016-June, IEEE, may 2016, pp. 536–543.

[42] A. Golovinskiy, V. G. Kim, and T. Funkhouser, *Shape-based recognition of 3d point clouds in urban environments*, in IEEE 12th International Conference on Computer Vision, 2009, sep 2009, pp. 2154–2161.

[43] I. Gordon and D. G. Lowe, *What and Where : 3D Object Recognition with Accurate Pose*, Toward Category-Level Object Recognition, (2006), pp. 67–82.

[44] N. Gorges, P. Fritz, and H. Woern, *Haptic Object Exploration Using Attention Cubes*, in Ki 2010: Advances in Artificial Intelligence, R. Dillmann, J. Beyerer, U. D. Hanebeck, and T. Schultz, eds., vol. 6359 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, jan 2010, pp. 349–357.

[45] N. Gorges, S. E. Navarro, D. Göger, and H. Wörn, *Haptic object recognition using passive joints and haptic key features*, in Proceedings - IEEE International Conference on Robotics and Automation, 2010, pp. 2349–2355.

[46] N. Gorges, S. E. Navarro, and H. Worn, *Haptic object recognition using statistical point cloud features*, in 2011 15th International Conference on Advanced Robotics (ICAR), 2011, pp. 15–20.

[47] K. Grauman and T. Darrell, *The pyramid match kernel: Discriminative classification with sets of image features*, in Proceedings of the IEEE International Conference on Computer Vision, vol. II, 2005, pp. 1458–1465.

[48] H. Gu, Y. Zhang, S. Fan, M. Jin, H. Zong, and H. Liu, *Model recovery of unknown objects from discrete tactile points*, in IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM, vol. 2016-Septe, IEEE, jul 2016, pp. 1121–1126.

[49] P. Guler, Y. Bekiroglu, X. Gratal, K. Pauwels, and D. Kragic, *What's in the container? Classifying object contents from vision and touch*, in IEEE International Conference on Intelligent Robots and Systems, sep 2014, pp. 3961–3968.

[50] K. He, X. Zhang, S. Ren, and J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, in 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, dec 2015, pp. 1026–1034.

[51] P. Hebert, N. Hudson, J. Ma, and J. Burdick, *Fusion of stereo vision, force-torque, and joint sensors for estimation of in-hand object location*, in Proceedings - IEEE International Conference on Robotics and Automation, 2011, pp. 5935–5941.

[52] G. Hetzel, B. Leibe, P. Levi, and B. Schiele, *3D object recognition from range images using local feature histograms*, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2001, pp. II–394.

[53] J. Hoelscher, J. Peters, and T. Hermans, *Evaluation of tactile feature extraction for interactive object recognition*, in IEEE-RAS International Conference on Humanoid Robots, vol. 2015-Decem, 2015, pp. 310–317.

[54] J. Ilonen, J. Bohg, and V. Kyrki, *Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing*, The International Journal of Robotics Research, 33 (2014), pp. 321–341.

[55] N. Jamali and C. Sammut, *Majority voting: Material classification by tactile sensing using surface texture*, IEEE Transactions on Robotics, 27 (2011), pp. 508–521.

[56] L. P. Jentoft, Y. Tenzer, D. Vogt, R. J. Wood, and R. D. Howe, *Flexible, stretchable tactile arrays from MEMS barometers*, in 2013 16th International Conference on Advanced Robotics (ICAR), IEEE, nov 2013, pp. 1–6.

[57] M. Jin, H. Gu, S. Fan, Y. Zhang, and H. Liu, *Object shape recognition approach for sparse point clouds from tactile exploration*, in 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO), no. December, dec 2013, pp. 558–562.

[58] M. K. Johnson, F. Cole, A. Raj, and E. H. Adelson, *Microgeometry capture using an elastomeric sensor*, in ACM SIGGRAPH 2011 Papers, SIGGRAPH '11, New York, NY, USA, 2011, ACM, pp. 46:1–46:8.

[59] M. Johnsson and C. Balkenius, *Sense of touch in robots with self-organizing maps*, IEEE Transactions on Robotics, 27 (2011), pp. 498–507.

[60] K. Kamiyama, H. Kajimoto, N. Kawakami, and S. Tachi, *Evaluation of a vision-based tactile sensor*, IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, 2 (2004), pp. 1542–1547.

[61] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, *Multisensor data fusion: A review of the state-of-the-art*, 2013.

[62] J. K. Kim, J. W. Wee, and C. H. Lee, *Sensor fusion system for improving the recognition of 3D object*, in IEEE Conference on Cybernetics and Intelligent Systems, 2004., vol. 2, 2004, pp. 1207–1212.

[63] E. Knoop and J. Rossiter, *Dual-mode compliant optical tactile sensor*, in 2013 IEEE International Conference on Robotics and Automation, May 2013, pp. 1006–1011.

[64] T. Kohonen, *Self-organized formation of topologically correct feature maps*, Biological Cybernetics, 43 (1982), pp. 59–69.

[65] O. Kroemer, C. H. Lampert, and J. Peters, *Learning dynamic tactile sensing with robust vision-based training*, IEEE Transactions on Robotics, 27 (2011), pp. 545–557.

[66] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, The Annals of Mathematical Statistics, 22 (1951), pp. 79–86.

[67] S. Lacey, C. Campbell, and K. Sathian, *Vision and touch: Multiple or multisensory representations of objects?*, Perception, 36 (2007), pp. 1513–1521.

[68] S. Lacey and K. Sathian, *Visuo-haptic multisensory object recognition, categorization, and representation*, Frontiers in Psychology, 5 (2014), p. 730.

[69] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning.*, Nature, 521 (2015), pp. 436–44.

[70] S. J. Lederman and R. L. Klatzky, *Hand movements: A window into haptic object recognition*, Cognitive Psychology, 19 (1987), pp. 342–368.

[71] N. F. Lepora and B. Ward-Cherrier, *Superresolution with an optical tactile sensor*, in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sept 2015, pp. 2686–2691.

[72] ——, *Tactile quality control with biomimetic active touch*, IEEE Robotics and Automation Letters, 1 (2016), pp. 646–652.

[73] H. Liu, D. Guo, and F. Sun, *Object Recognition Using Tactile Measurements: Kernel Sparse Coding Methods*, IEEE Transactions on Instrumentation and Measurement, 65 (2016), pp. 656–665.

[74] H. Liu, X. Song, J. Bimbo, L. Seneviratne, and K. Althoefer, *Surface material recognition through haptic exploration using an intelligent contact sensing finger*, in IEEE International Conference on Intelligent Robots and Systems, 2012, pp. 52–57.

[75] H. Liu, Y. Yu, F. Sun, and J. Gu, *Visual Tactile Fusion for Object Recognition*, IEEE Trans. on Automation Science and Engineering, (2016), pp. 1–13.

[76] S. Liu and W. Deng, *Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size*, in 2015 3rd IAPR Asian Conference on Pattern Recognition, IEEE, nov 2015, pp. 730–734.

[77] D. G. Lowe, *Object recognition from local scale-invariant features*, in Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, 1999, pp. 1150–1157.

[78] S. Luo, X. Liu, K. Althoefer, and H. Liu, *Tactile object recognition with semi-supervised learning*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9245, 2015, pp. 15–26.

[79] S. Luo, W. Mou, K. Althoefer, and H. Liu, *Iterative Closest Labeled Point for Tactile Object Shape Recognition*, in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, oct 2016, pp. 3137–3142.

[80] S. Luo, W. Mou, M. Li, K. Althoefer, and H. Liu, *Rotation and Translation Invariant Object Recognition with a Tactile Sensor*, in IEEE Sensors Conference, vol. 2014-Decem, nov 2014, pp. 1030–1033.

[81] M. Madry, L. Bo, D. Kragic, and D. Fox, *ST-HMP: Unsupervised Spatio-Temporal feature learning for tactile data*, in Proceedings - IEEE International Conference on Robotics and Automation, 2014, pp. 2262–2269.

[82] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.

[83] M. Meier, M. Schöpfer, R. Haschke, and H. Ritter, *A probabilistic approach to tactile shape reconstruction*, IEEE Transactions on Robotics, 27 (2011), pp. 630–635.

[84] J. L. Mundy, *Object Recognition in the Geometric Era: A Retrospective*, Springer Berlin Heidelberg, 2006, pp. 3–28.

[85] S. E. Navarro, N. Gorges, H. Wörn, J. Schill, T. Asfour, and R. Dillmann, *Haptic object recognition for multi-fingered robot hands*, in Haptics Symposium 2012, HAPTICS 2012 - Proceedings, 2012, pp. 497–502.

[86] F. N. Newell, M. O. Ernst, B. S. Tjan, and H. H. Bu, *Research Article VIEWPOINT DEPENDENCE IN VISUAL AND HAPTIC*, Psychological Science, 12 (2001), pp. 37–42.

[87] F. N. Newell, A. T. Woods, M. Mernagh, and H. H. Bülthoff, *Visual, haptic and crossmodal recognition of scenes*, Experimental Brain Research, 161 (2005), pp. 233–242.

[88] S. Omata, Y. Murayama, and C. E. Constantinou, *Real time robotic tactile sensor system for the determination of the physical properties of biomaterials*, Sensors and Actuators, A: Physical, 112 (2004), pp. 278–285.

[89] A. Petrovskaya and O. Khatib, *Global localization of objects via touch*, IEEE Transactions on Robotics, 27 (2011), pp. 569–585.

[90] A. Petrovskaya, O. Khatib, S. Thrun, and A. Y. Ng, *Bayesian estimation for autonomous object manipulation based on tactile sensors*, in Proceedings - IEEE International Conference on Robotics and Automation, vol. 2006, may 2006, pp. 707–714.

[91] Z. Pezzementi, E. Plaku, C. Reyda, and G. D. Hager, *Tactile-object recognition from appearance information*, IEEE Transactions on Robotics, 27 (2011), pp. 473–487.

[92] J. Platt, *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*, Advances in large margin classifiers, 10 (1999), pp. 61–74.

[93] M. Prats, P. J. Sanz, and A. P. Del Pobil, *Reliable non-prehensile door opening through the combination of vision, tactile and force feedback*, Autonomous Robots, 29 (2010), pp. 201–218.

[94] J. A. Pruszynski and R. S. Johansson, *Edge-orientation processing in first-order tactile neurons.*, Nature Neuroscience, 17 (2014), pp. 1404–1409.

[95] N. I. Rafla, *Visually guided tactile and force-torque sensing for object recognition and localization*, PhD thesis, 1991.

[96] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.*, vol. 14, 2004.

[97] S. Ratnasingam and T. M. McGinnity, *A comparison of encoding schemes for haptic object recognition using a biologically plausible spiking neural network*, in IEEE International Conference on Intelligent Robots and Systems, 2011, pp. 3446–3453.

[98] C. Roke, C. Melhuish, T. Pipe, D. Drury, and C. Chorley, *Deformation-based tactile feedback using a biologically inspired sensor and an improved display*, in Taros, R. Groß, L. Alboul, C. Melhuish, M. Witkowski, T. J. Prescott, and J. Penders, eds., vol. 6856 LNAI of Lecture Notes in Computer Science, Springer Berlin Heidelberg, jan 2011, pp. 114–124.

[99] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, *3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints*, mar 2006.

[100] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, *ImageNet Large Scale Visual Recognition Challenge*, International Journal of Computer Vision, 115 (2015), pp. 211–252.

[101] R. Russell, *Object recognition by a 'smart' tactile sensor*, in Proceedings of the Australian Conference on Robotics . . . , 2000, pp. 93–98.

[102] M. Sanchez-Fibla, A. Duff, and P. F. M. J. Verschure, *A sensorimotor account of visual and tactile integration for object categorization and grasping*, in Proceedings - IEEE International Conference on Robotics and Automation, 2013, pp. 107–112.

[103] J. Schmidhuber, *Deep Learning in neural networks: An overview*, Neural Networks, 61 (2015), pp. 85–117.

[104] A. Schmitz, M. Maggiali, L. Natale, B. Bonino, and G. Metta, *A tactile sensor for the fingertips of the humanoid robot icub*, in 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Oct 2010, pp. 2212–2217.

[105] A. Schneider, J. Sturm, C. Stachniss, M. Reisert, H. Burkhardt, and W. Burgard, *Object identification with*

tactile sensors using bag-of-features, in Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on DOI - 10.1109/IROS.2009.5354648, 2009, pp. 243–248.

[106] M. Schöpfer, H. Ritter, and G. Heidemann, *Acquisition and application of a tactile database*, in Proceedings - IEEE International Conference on Robotics and Automation, 2007, pp. 1517–1522.

[107] J. Sinapov, V. Sukhoy, R. Sahai, and A. Stoytchev, *Vibrotactile recognition and categorization of surfaces by a humanoid robot*, IEEE Transactions on Robotics, 27 (2011), pp. 488–497.

[108] H. Soh and Y. Demiris, *Incrementally learning objects by touch: Online discriminative and generative models for tactile-based recognition*, IEEE Transactions on Haptics, 7 (2014), pp. 512–525.

[109] H. Soh, Y. Su, and Y. Demiris, *Online spatio-temporal Gaussian process experts with application to tactile classification*, in IEEE International Conference on Intelligent Robots and Systems, 2012, pp. 4489–4496.

[110] A. J. Spiers, M. V. Liarokapis, B. Calli, and A. M. Dollar, *Single-Grasp Object Classification and Feature Extraction with Simple Robot Hands and Tactile Sensors*, IEEE Transactions on Haptics, 9 (2016), pp. 207–220.

[111] S. Stassi, V. Cauda, G. Canavese, and C. F. Pirri, *Flexible tactile sensing based on piezoresistive composites: A review*, 2014.

[112] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, *Multimodal Feature-based Surface Material Classification*, IEEE Transactions on Haptics, (2016), pp. 1–1.

[113] F. Sun, C. Liu, W. Huang, and J. Zhang, *Object Classification and Grasp Planning Using Visual and Tactile Sensing*, IEEE Transactions on Systems, Man, and Cybernetics: Systems, (2016), pp. 1–11.

[114] R. Szeliski, *Computer Vision : Algorithms and Applications*, Computer, 5 (2010), p. 832.

[115] D. Taddeucci, C. Laschi, R. Lazzarini, R. Magni, P. Dario, and A. Starita, *An approach to integrated tactile perception*, in Proceedings

of International Conference on Robotics and Automation, vol. 4, 1997, pp. 3100–3105.

[116] S. Takamuku, A. Fukuda, and K. Hosoda, *Repetitive grasping with anthropomorphic skin-covered hand enables robust haptic recognition*, in 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2008, pp. 3212–3217.

[117] Y. Tenzer, L. Jentoft, and R. Howe, * Inexpensive and Easily Customized Tactile Array Sensors using MEMS Barometers Chips *, IEEE R&A Magazine, 21 (2012), p. 2013.

[118] L. R. Tucker, *An inter-battery method of factor analysis*, Psychometrika, 23 (1958), pp. 111–136.

[119] N. Vasconcelos, J. Pantoja, H. Belchior, F. V. Caixeta, J. Faber, M. A. M. Freire, V. R. Cota, E. Anibal de Macedo, D. a. Laplagne, H. M. Gomes, and S. Ribeiro, *Cross-modal responses in the primary visual cortex encode complex objects and correlate with tactile discrimination.*, Proceedings of the National Academy of Sciences of the United States of America, 108 (2011), pp. 15408–15413.

[120] A. Vedaldi and K. Lenc, *MatConvNet*, in Proceedings of the 23rd ACM international conference on Multimedia - MM '15, New York, New York, USA, 2015, ACM Press, pp. 689–692.

[121] G. Vezzani, N. Jamali, U. Pattacini, G. Battistelli, L. Chisci, and L. Natale, *A novel Bayesian filtering approach to tactile object recognition*, in IEEE-RAS International Conference on Humanoid Robots, IEEE, nov 2016, pp. 256–263.

[122] A. Wald, *Sequential tests of statistical hypotheses*, The Annals of Mathematical Statistics, 16 (1945), pp. 117–186.

[123] K. Weiss and H. Worn, *The working principle of resistive tactile sensor cells*, in IEEE International Conference Mechatronics and Automation, 2005, vol. 1, 2005, pp. 471–476.

[124] N. Wettels, V. J. Santos, R. S. Johansson, and G. E. Loeb, *Biomimetic Tactile Sensor Array*, Advanced Robotics, 22 (2008), pp. 829–849.

[125] B. Winstone, G. Griffiths, C. Melhuish, T. Pipe, and J. Rossiter, *TACTIP - Tactile fingertip device, challenges in reduction of size to ready for robot hand integration*, in 2012 IEEE International Conference on Robotics and Biomimetics, ROBIO 2012 - Conference Digest, 2012, pp. 160–166.

[126] B. Winstone, G. Griffiths, T. Pipe, C. Melhuish, and J. Rossiter, *TACTIP - Tactile fingertip device, texture analysis through optical tracking of skin features*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8064 LNAI, 2013, pp. 323–334.

[127] J. Yang, H. Liu, F. Sun, and M. Gao, *Object recognition using tactile and image information*, in 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE, dec 2015, pp. 1746–1751.

[128] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, *Measurement of shear and slip with a gelsight tactile sensor*, in 2015 IEEE International Conference on Robotics and Automation (ICRA), May 2015, pp. 304–311.

[129] W. Yuan, S. Wang, S. Dong, and E. Adelson, *Connecting look and feel: Associating the visual and tactile properties of physical materials*, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 4494–4502.

[130] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, *Shape-independent hardness estimation using deep learning and a gelsight tactile sensor*, in 2017 IEEE International Conference on Robotics and Automation (ICRA), May 2017, pp. 951–958.

[131] M. D. Zeiler, *ADADELTA: An Adaptive Learning Rate Method*, arXiv, (2012), p. 6.

[132] F. Zernike, *Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode*, Physica, 1 (1934), pp. 689–704.

[133] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, *Local Features and Kernels for Classi cation of Texture and Object Categories: A*

Comprehensive Study, in Computer Vision and Pattern Recognition Workshop, 2006, pp. 13–13.

[134] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Ozer, and E. Steinbach, *Deep Learning for Surface Material Classification Using Haptic and Visual Information*, IEEE Transactions on Multimedia, 18 (2016), pp. 2407–2416.